UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E INVESTIGAÇÃO OPERACIONAL

Modelling Influenza Strain Competition Dynamics and Transmission Fitness

Afonso Dimas Martins

Mestrado em Bioestatística

Dissertação orientada por:
Dr. Erida Gjini
Prof. Maria Helena Mouriño Nunes

2019

Acknowledgements

I would like to thank my friends and colleagues at Instituto Gulbenkian de Ciência (IGC). I was for almost two years in possibly the best work environment, surrounded by welcoming and inspiring people.

I am grateful to the organizers of the Mathematical Biology on the Mediterranean Conference for giving me the opportunity to be part of a Summer School intended for PhD students and to present my work informally to an audience specialized in mathematical modelling. I learned a lot, made friends from distant points of Europe and had once-in-a-lifetime adventures in the beautiful greek island of Samos.

I also want to thank Professor Maria Helena Mouriño Nunes for the comments on the statistical part of our modelling work. Her feedback and outside perspective were very helpful.

Last, but certainly not least, I must thank my supervisor Dr. Erida Gjini. She guided me through a completely new field, spending many hours and exchanging many e-mails with thorough discussions and critical questions that forced me to push further. Her support and enthusiasm didn't seem to fade, from the beginning of my internship to last stretch of the thesis. I'm certain I want to pursue research in mathematical modelling because of her influence. I'm forever grateful.

Resumo

A estimação de diferenças de aptidão (ou *fitness*) entre indivíduos, particularmente de espécies microbiais patogénicas, é uma área muito ativa de investigação. Nesta área, dado o aumento constante da evolução de resistência a drogas, mecanismos de evasão de vacinas e emergência viral, tem sido crucial entender as diferenças de aptidão viral. Recentemente, têm sido propostas várias abordagens experimentais e de modelação com o objetivo de quantificar a diversidade viral. Isto é particularmente importante quando se pretende comparar patogénios resistentes ou sensíveis a tratamentos, e quando se procura planear medidas de contenção e prevenção. Esta tese é motivada por uma série de experiências de transmissão de duas estirpes de *influenza*, o vírus responsável pela gripe. A gripe é uma doença infeciosa que afeta populações animais e humanas, e é uma causa significativa de morbilidade e mortalidade no mundo. Estima-se que ocorram até 650 000 mortes durante eventos epidémicos anuais. O vírus é transmitido através de aerossóis espalhados por espirros ou tosse. É comum a infeção estar associada a complicações causadas por outros agentes, como a coinfecção com a bactéria que causa a pneumonia. A prevenção de epidemias gripais graves é feita recorrendo à ação de antivirais, no entanto, a resistência aos mesmos está a aumentar. Através de alterações genómicas, novas estirpes do vírus podem emergir, havendo o risco de algumas serem resistentes ao reconhecimento do sistema imunitário ou dos antivirais. No caso de a resistência não acarretar custos de fitness de transmissão, existe o potencial de causar um evento pandémico.

A estimação de aptidão viral é usualmente baseada em métodos estatísticos que comparam o fitness replicativo relativo de duas estirpes, em culturas de células, tecidos ou hospedeiros individuais. De modo a associar a aptidão num hospedeiro com a aptidão de transmissão entre hospedeiros, uma abordagem experimental, conhecida como misturas competitivas, foi proposta por Hurt et al. (2010). Estas experiências envolveram a infeção de furões com uma mistura de uma estirpe de gripe suscetível a um antiviral comum e uma estirpe resistente ao mesmo, e a subsequente medição diária das proporções relativas destas estirpes de modo a investigar se um vírus se está a replicar mais rapidamente que o outro. Os dados obtidos neste estudo

servem como ponto de partida para este trabalho. Para quantificar as diferenças entre estirpes suscetíveis e estirpes resistentes, foi proposto mais tarde por McCaw et al (2011) um modelo matemático que traduz essas diferenças em termos de um coeficiente baseado nas suas taxas de crescimento. Esse modelo é simplista e prevê apenas cenários em que uma estirpe leva a outra estirpe à extinção. Estes autores estimaram uma ligeira vantagem da estirpe resistente, isto é, tem uma taxa de crescimento maior que a estirpe suscetível. No entanto, os dados têm uma grande variabilidade, indicando que provavelmente a estirpe resistente não conduz sempre a estirpe suscetível à extinção. Além disso, este modelo não é capaz de explicar a coexistência de ambas as estirpes. Nem todos os vírus são capazes de se transmitir de um hospedeiro para outro, um conceito denominado de *bottleneck* de transmissão, refletido pelo parâmetro N, o número total de vírus que se transmitem, independentemente da estirpe. Esse modelo prevê um *bottleneck* estreito, isto é, poucas partículas virais no total são transmitidas, $N \approx 4$. Isto permite explicar a grande variabilidade observada nas proporções relativas das duas estirpes, no entanto é inconsistente com estimativas atuais para o número de vírus transmitidos entre hospedeiros. A heterogeneidade num hospedeiro, isto é, a co-circulação de diferentes estirpes de gripe, é comum, e resulta em competição pelos recursos e espaço do hospedeiro. Dadas estas preocupações, é de uma grande relevância modelar e ganhar conhecimento mais aprofundado das dinâmicas de competição entre estirpes de gripe e como isso afeta as suas capacidades de transmissão entre hospedeiros. O principal objetivo desta tese é aplicar um modelo alternativo a estes dados de transmissão de misturas de estirpes que permita explicar os dados recorrendo às dinâmicas de crescimento e de competição, e com mais flexibilidade para estimar o número de vírus transmitidos.

Nesta tese, apresentamos um modelo matemático baseado nas dinâmicas de competição entre estirpes de gripes suscetíveis e resistentes a antivirais. O nosso modelo, baseado nas equações de competição de Lotka-Volterra, é aplicado aos dados experimentais de misturas de estirpes gripais, com o objetivo de compreender como os mecanismos de competição intra- e inter-estirpe afetam a aptidão relativa de transmissão entre hospedeiros. No Capítulo 2 introduzimos o modelo e ilustramos as suas previsões num hospedeiro. Aí mostramos como este modelo, ao contrário de abordagens clássicas baseadas apenas num coeficiente de aptidão, não está limitado a cenários de exclusão competitiva, isto é, a estirpe resistente leva a suscetível à extinção ou vice-versa. Dois novos cenários ecológicos emergem: coexistência estável de ambas as estirpes, e um cenário de bi-estabilidade, no qual, dependendo das condições iniciais, o sistema colapsa para um dos cenários de exclusão competitiva. No Capítulo 3 fazemos a ponte entre as dinâmicas num hospedeiro e as subsequentes dinâmicas de transmissão entre hospedeiros. Também validamos as

previsões do modelo com recurso a simulações. No Capítulo 4 o modelo é ajustado aos dados de transmissão de misturas de estirpes de influenza (McCaw et al (2011)) através de optimização em R. A incerteza à volta das estimativas do modelo é quantificada recorrendo a simulações. Estas simulações baseiam-se em gerar dados artificiais equivalentes aos dados originais, em que simulam a variabilidade observada causada por erro experimental (gerando dados de forma uniforme em redor das observações) ou pelo efeito de *bottleneck* (gerando dados através de reamostragem das condições iniciais com um modelo Binomial). Este último método tem a vantagem de não só permitir quantificar a incerteza dos parâmetros do modelo, como também obter uma estimativa do número total de vírus transmitidos entre hospedeiros, N.

A nossa proposta é capaz de inferir com precisão parâmetros de crescimento e de competição, e prevê neste contexto específico um cenário de coexistência das duas estirpes virais, isto é, que a estirpe suscetível e a resistente são transmitidas em conjunto, com base nestes dados de experiências de misturas competitivas. O nosso estudo tem implicações para a epidemiologia e modelação matemática, e aprofunda o conhecimento dos resultados experimentais de misturas competitivas no geral. Ao permitir a possibilidade de competição dependente da frequência e hierarquias entre estirpes, este modelo expande o alcance de cenários ecológicos que podem ser capturados, incluindo coexistência e bi-estabilidade, antes da ativação do sistema imune. A coexistência mútua resultante das dinâmicas de competição pode ajudar a explicar a heterogeneidade viral observada ao nível populacional. Adicionalmente, o modelo prevê um número de vírus transmitidos, $N \approx 230$, compatível com a literatura recente de *influenza* em humanos. Quanto mais flexível um modelo é para capturar não-linearidade nos dados, menos hipóteses existem para atribuir flutuações observadas nos dados a pura variabilidade causada por um *bottleneck* estreito. Esta tese deve servir como *proof-of-concept*, sendo, no entanto, a abordagem geral o suficiente para ser aplicada a cenários ecológicos entre outras estirpes ou outras espécies que compitam entre si.

Keywords: *influenza*, dinâmicas virais dentro-de-hospedeiro, *fitness* de transmissão, modelação matemática, equações diferenciais ordinárias

Abstract

There have been proposed in the recent years many experimental and modelling approaches to quantify viral diversity. This is particularly important when comparing drug-resistant with drug-sensitive pathogens and when designing control measures.

We present a general mathematical and statistical framework that focuses on the competition dynamics between two strains of influenza. Managing influenza outbreaks has been done extensively over the years using antivirals, however resistance is on the rise. Via mutational changes, new strains of virus can emerge that are resistant to these antivirals. It is important to quantify fitness differences of such mutants with wild-type strains, to predict epidemiological outcomes and design control measures. The resistance to antivirals can sometimes carry no cost of transmission fitness, having the potential to cause a pandemic event. Given these concerns, it is of major relevance to model and gain an understanding of the dynamics of competing influenza strains within host, and how this affects their relative transmissibility between hosts.

Our model is based on the Lotka-Volterra equations and is applied to data from competitive mixture experiments, with the aim is to comprehend how the intra- and inter-strain competition affects the relative strain transmission between hosts. The model is validated through a simulation approach and the parameter uncertainty is quantified using observations from the simulation procedure. Our framework can accurately infer parameter values and, for this data, predicts a scenario of coexistence of the antiviral susceptible and antiviral resistant strains. Additionally, we predict, compared with previous estimates, a relatively higher transmission bottleneck size, i.e. a total number of virions transmitted between hosts of approximately 230. This thesis serves as a proof-of-principle, with the model being general enough to be applied to a variety of ecological interactions that involve competition and allow for results beyond competitive exclusion.

Keywords: influenza, within-host viral dynamics, transmission fitness, mathematical modelling, ordinary differential equations

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Species interact in a myriad of ways. These interactions shape directly or indirectly the structure of the communities in which they are inserted (Wootton and Emmerson, 2005). The nature of these interactions can vary depending on the ecological and evolutionary context in which they occur, which makes it difficult to define and measure. For simplicity, we will here define only a few simple concepts and categories. The interactions between organisms can be classified as intra- or inter-specific, if they occur within individuals of the same species or between different species, respectively. In an ecological community, these interactions act at distinct levels and affect the individuals involved in different ways. In mutualism, all participating species derive a benefit; in commensalism, one species drives the benefit while the other is unaffected; in competition, all the species involved are harmed (Martin and Schwab, 2013). These are just a few examples of many possible interactions that can be established between organisms. Competition will be the main focus of this thesis.

The overwhelming complexity of relations and interactions between species is often studied with the use of models. Models are a simplification of reality. Through formal representation it is possible to summarize the key mechanisms of a dynamical system. The literature of ecology and evolution is replete with mathematical models that can help us predict and give insights about biological phenomena (Gillman, 2009). Likewise, many theoretical models have been proposed to study pathogen dynamics, evolutionary patterns and epidemiological consequences.

In this thesis, we present a mathematical and statistical framework of viral competition and apply it to influenza data to gain insights into its transmission. This introduction outlines the motivation to model influenza dynamics, the state-of-the-art and our proposed objectives.

1.1 The influenza virus

Influenza is an infectious disease that greatly affects animal and human populations. This is caused by the influenza virus, transmitted via host-to-host contact, and is a major cause of morbidity and mortality around the world. During annual epidemic events, it is estimated 3 to 5 million severe cases and 290 000 to 650 000 respiratory deaths (WHO). The main victims are children, being 20% affected worldwide each year, and people suffering concomitant diseases, such as chronic respiratory disease or diabetes (Turner et al., 2003). It causes economic losses associated with hospitalization costs, vaccination and public health actions such as quarantines. The infection may range from asymptomatic to more severe respiratory syndromes and be associated with a secondary infection such as with the bacteria *Streptococcus pneumoniae* (Nicholson, 1992). Usually, the virus is transmitted through respiratory droplets discharged by sneezing or coughing (Nicholson, 1992).

Any virus can only replicate itself by infecting a host cell and using its molecular machinery to reproduce. The influenza virus is no exception to this rule. In this process two proteins on the viral surface have a leading role: haemagglutinin (HA), which is responsible for the correct binding to the target cell, thereby assisting on the entry of the viral genome on the cell, and neuraminidase (NA), in charge of cleaving sugars that bind the mature viral particle, in this way guarantying a correct exit of the viral progeny from the cell (Suzuki, 2005). The antigenic properties of the virus depend on HA and NA, leading to a large diversity of different strains. Due to their high importance in the viral replication cycle, these proteins have also become major targets for antiviral treatment (Wilson and Itzstein, 2003).

New variants can emerge due to two evolutionary mechanisms that enable the virus to alter its surface proteins: antigenic drift (Figure 1.1 a)) and antigenic shift (Figure 1.1 b)). Through antigenic drift, the virus accumulates mutations on those proteins, eventually originating a strain that is able to evade recognition by the immune system or antiviral action (Potter, 2001). Through antigenic shift, a strain combines with one or more different strains, leading to new viral subtypes (Webster, 1999). These mechanisms have the chance to give rise to a variant strain that is resistant to antivirals. If these new strains are highly transmissible they could give rise to a pandemic event (Webster, 1999; Parrish and Kawaoka, 2005). This is why understanding fitness differences between existing and new strains is important.

When studying influenza, some concepts are defined to measure the success of particular strains. One is replicative fitness, which is how much genetic material is passed on to the next generation (Domingo, 2010). Commonly, it is examined by growing viral strains in separate cell

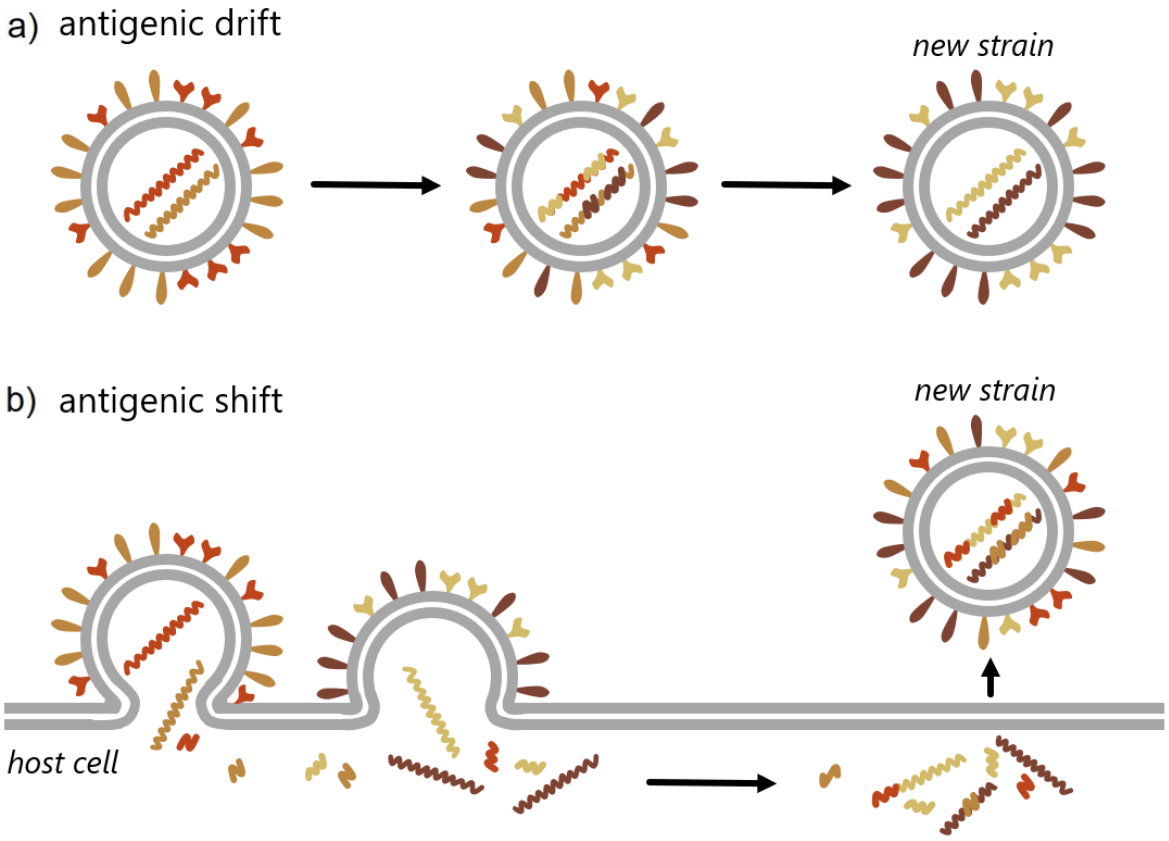

Figure 1.1: Two mechanisms of viral evolution. a) In antigenic drift, a virus accumulates mutations that change its surface proteins. In the right conditions of selective pressure a new strain may be originated. b) In antigenic shift, if different strains of viruses infect the same host cell, there's a chance of creating a *de novo* strain when the progeny is exiting the cell.

cultures, and comparing the rates of growth (Hurt et al., 2010). The other is transmission fitness, which is how well a virus transmits from a donor to a recipient host (Wargo and Kurath, 2012). It is usually studied by exposing naïve hosts to infected hosts, and analysing how much virions were transmitted (Duan et al., 2010). Also, the transmission bottleneck expresses the amount of viral particles that are transmitted between hosts (Leonard et al., 2017), and its size affects the viral diversity (Leonard et al., 2017; Gutiérrez et al., 2012), as is illustrated in Figure 1.2. These measures of fitness inform us how prevalent a viral strain could be at the epidemiological level.

So, given the global consequences of this virus, a deep knowledge of influenza is a key step towards the establishment of improved prevention, control and treatment strategies. Thus, as we will describe in the next section, mathematical modeling has risen as a tool to quantify and study influenza infection.

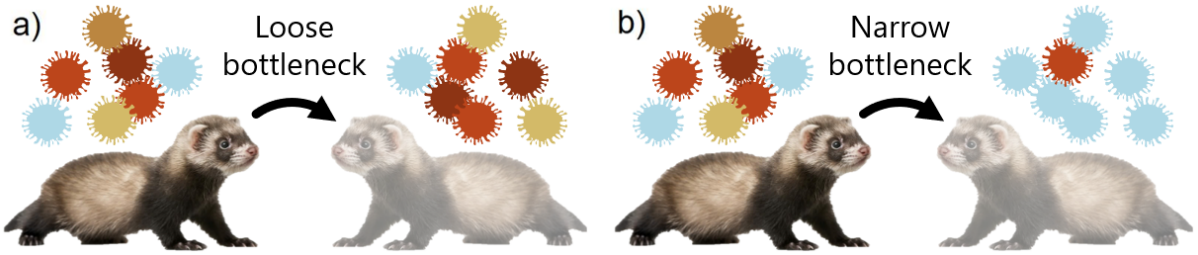

Figure 1.2: Illustration of the impact of the bottleneck size on the viral diversity that initiates an infection in a new host. a) In a loose bottleneck, many virions are transmitted between hosts, allowing the preservation of the relative strain proportions. b) In a narrow bottleneck, few viral particles initiate infection in the new host, so stochastic fluctuations play a big role and genetic viral diversity is rarely preserved.

1.1.1 State-of-the-art of mathematical modelling of influenza

Due to the global threat of influenza, many mathematical models have been proposed over the years in order to better understand how this virus behaves within-host and how it spreads and affects the population. All models have their advantages and limitations, and focus on different aspects of an influenza infection.

Most models have been aimed at the epidemiological level (Beauchemin and Handel, 2011). Based in dividing the population in compartments of susceptibles, infected and recovered - the classical SIR model (Mikolajczyk et al., 2009; Coburn et al., 2009) - they can be altered to include more complex dynamics such as disease resistance (Khanh, 2016) or vaccination (Feng et al., 2011). This class of models is centered on the big-picture of the influenza infection and is usually applied to inform public health authorities to plan for contingencies to contain epidemics (Saunders-Hastings et al., 2017; McVernon et al., 2007; Lee et al., 2013; Ferguson et al., 2005; Germann et al., 2006).

Models at the within-host level have also been applied thoroughly in the literature (Boianelli et al., 2015). Most of these adopt the target-cell model, which consists of uninfected cells, infected cells and virions (Baccam et al., 2006). In a similar way to the SIR models, they can be modified to include more levels of detail. Some authors have extended this framework to include the effect of a latent phase (Holder and Beauchemin, 2011), immune response (Bocharov and Romanyukha, 1994; Hancioglu et al., 2007) and even viral-viral (Smith, 2018) or viral-bacterial co-infections (Smith, 2018; Cheng et al., 2017).

The models can be either deterministic (Baccam et al., 2006; Bocharov and Romanyukha, 1994) or stochastic (Ferguson et al., 2005; Germann et al., 2006; Xu et al., 2007). In deterministic models, if provided the same parameter values and same initial conditions, the same output will be always achieved (Renard et al., 2013). In a stochastic model however, the system is described

with implicit random processes, so its solutions are not unique, i.e. the same parameters and initial conditions will always lead to different outputs (Renard et al., 2013). Stochastic models are more realistic than the deterministic since they account for the inherent uncertainty of biological processes, especially for small populations or early stages of an epidemic. Deterministic and stochastic models, however, should not be seen as opposing strategies, but rather as complementary approaches (Britton, 2010).

In the context of this thesis, as will be described in detail in Chapter 2, we construct a deterministic within-host model for influenza co-infection. Yet in our approach we apply stochastic steps to account for uncertainty during transmission, bridging, in this way, with between-host dynamics.

Next we introduce the experimental setup and the later proposed mathematical and statistical modelling approaches that motivated this thesis.

1.1.2 Influenza relative fitness estimation with competitive mixtures experimental model

In order to study the relative fitness of two strains of influenza, Hurt et al. (2010) designed a novel experimental model using ferrets co-infected with two strains - an experimental model they called "competitive-mixtures". The strains in study were a mutant strain (H274Y MUT), that is resistant to the antiviral oseltamivir and a wild-type strain (H274 WT), sensitive to oseltamivir. Naïve ferrets were infected either with a pure population of one strain or co-infected with mixtures of both strains in differing ratios. In short, two groups of ferrets, donor 1 (D1) and donor 2 (D2), were infected with the two strains at different proportions (0% MUT-100% WT, 20% MUT-80% WT, etc.). These were then put in contact with five recipient ferrets (R1), which were analysed daily until one of the strains could be detected. They then acted as donor after being put in contact with a new generation of naïve ferrets (R2). An example of these transmission experiments is represented in Figure 1.3. Note that the experimental framework by Hurt et al. (2010) was more complex and involved other transmission events from donor to recipient 1 and recipient 2, but here we focus only on a primary transmission event where the proportion in the donor and recipient is known.

Therefore, there were a total of 9 transmission events (one repeated observation was removed), which are represented in Figure 1.4 a). This study stands out from most studies that intend to study viral fitness differences by using a strain mixture instead of pure population infection. In this experimental model, a fitness advantage of a strain is deduced from an observed

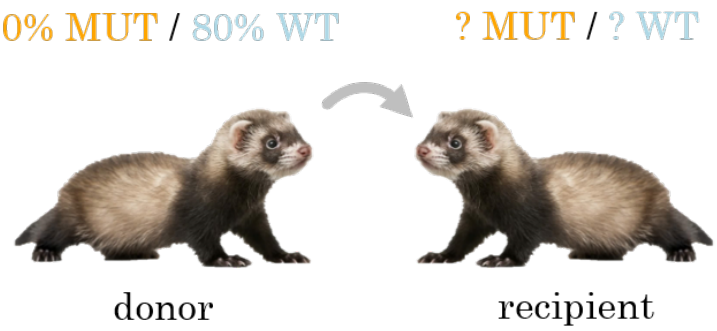

Figure 1.3: Illustration of a transmission experiment of competitive-mixtures of influenza strains. Naïve ferret (donor) are inoculated with known combinations of the two influenza strains and then put in contact with new ferrets (recipient). The proportion of the two strains in the recipient ferrets is measured and then plotted for all pairs as seen in Figure 1.4 a).

increase in the replication rate. This means that, in this case, the replication fitness is directly related to the transmission fitness. A strain with a higher replication rate would produce a larger progeny, having in this way a higher proportion of cells that could transmit virions. Nonetheless, this may not be strictly true. As it will be discussed in more detail in Section 1.1.4, a strain could have a high replication fitness but a low transmission fitness, being well adapted to one host but incapable of successful infection in another (Rodpothong and Auewarakul, 2012).

The main focus of these experiments was quantifying the within-host fitness differences between the two strains, by analysing the entire time-series within a recipient ferret, and not study the relative transmission differences. However the authors briefly hypothesize that the relative transmission fitness is equal to 1, i.e. both the mutants and the wild-type strains have equal transmissibility between hosts. They refer this because about as many points are above and below the diagonal of the plot, however this is only a qualitative assessment based on those results.

The results of this study could be extended to comparable seasonal mutant strains. They argue that the rising of the H274Y H1N1 strain could be due to an equal or greater fitness level compared to its wild-type. The authors show that although the mutant strain has a replication fitness cost compared to the wild-type when present in competitive mixtures, it has about the same transmission fitness. This last conclusion was drawn exclusively from a qualitative analysis (see section 1.1.4 for more details). In order to get a proper quantification, a model was later proposed, as will be outlined in the following section.

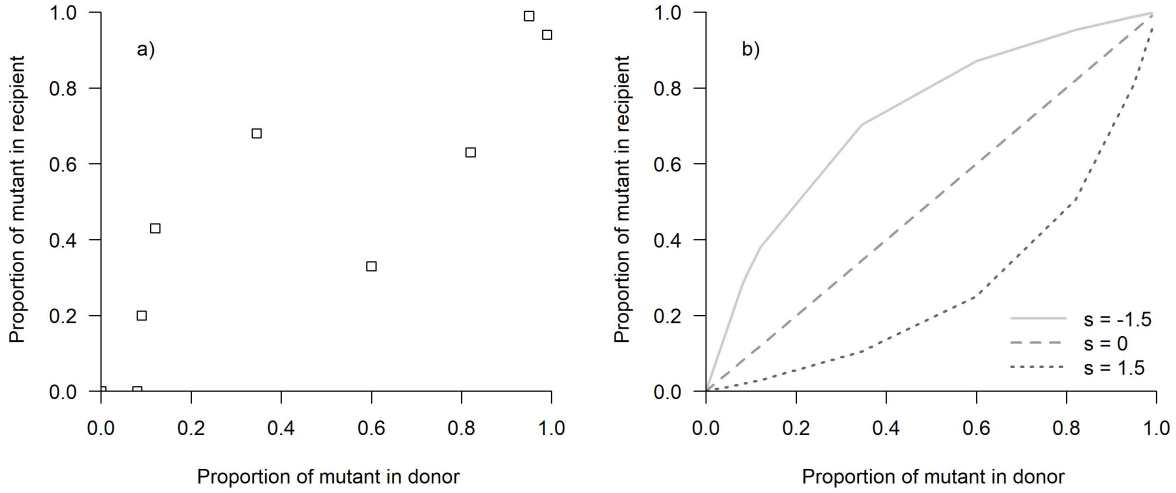

Figure 1.4: Competitive mixtures transmission data and data and illustration of the fitness coefficient s. a) Summary of the transmission events carried by Hurt et al. (Hurt et al., 2010). Each square corresponds to a separate transmission event between a donor and a recipient ferrets. The observed values of the proportion of the mutant in the donor do not correspond exactly to the inoculated values (0/100, 20/80, 50/50, etc.) due to intrinsic stochastic variability caused by viral dynamics in the donor. b) Illustration of the competitive exclusion scenarios obtained by changing the value of s. If s is negative (smooth line), the mutant strain will always be at a higher proportion in the recipient, while if s is positive (dotted line), the mutant will always be outcompeted by the wild-type. If $s = 0$ (dashed diagonal line), then there's no clear fitness advantage to either strain.

1.1.3 Influenza relative transmissibility assessment

As a follow-up from these experiments, a later theoretical study by McCaw et al. (2011) performed a quantitative assessment of the transmission fitness of the resistant and susceptible strains. They proposed a mathematical and statistical framework to capture the key characteristics of transmission of the experimental model of competitive mixtures. In this case, they did not model the entire time series but only the transmission events.

They derived a function that gives the proportion of mutant in the recipient, P, given the proportion of the mutant in the donor, p:

$$P(p, s) = \frac{p}{p + (1 - p)e^s}, \qquad -\infty < s < \infty \tag{1.1}$$

where s describes the relative viral fitness of the mutant strain compared with the wild-type, and accounts for: (1) secretion from the donor, (2) the initial extinction probability in the recipient and (3) subsequent growth in the recipient. As illustrated in Figure 1.4 b), if $s < 0$, the mutant has a fitness advantage compared to the wild-type and if $s > 0$, the mutant has a fitness disadvantage. From fitting this simple model to the transmission data, they estimated $s = -0.2597$, which means a slight fitness advantage to the mutant strain, however with a

very wide confidence interval for s $(-1.3509, 0.8329)$, and attributing a large role to inherent stochasticity. Furthermore, when estimating the bottleneck size with their model, they estimated a very small bottleneck size ($N_b = 3.8$), assigning the spread in the data to stochasticity, a finding that has been challenged as too low by other studies modeling influenza transmission in horses and pigs (Stack et al., 2013) and humans (Leonard et al., 2017; Poon et al., 2016).

1.1.4 Limitations of the McCaw et al. model

The model developed by McCaw et al. (2011) leaves out many components of this biological system that could prove to be informative and helpful in interpreting the result. The main limitation in assuming viral exponential growth is that such formulation does not allow for the possibility that the success of one strain may be frequency-dependent, thus preventing the possibility of y-observations crossing the diagonal at some mixture ratio (see Figure 4.1). By having no implicit or explicit interaction between strains, only two scenarios of competitive exclusion are possible in this model: either the mutant strain always wins (all points above the diagonal) or the wild-type strain outcompetes the other (all points below the diagonal). In order to capture the data points above and below the diagonal, this model also had to assume high stochasticity. This results in a very wide confidence interval for s, inevitably estimating as well a very low bottleneck size, responsible for such variability. By allowing only two scenarios, there is an unavoidable tendency of this formulation to bias estimates of bottleneck size to very low values and assign a big role to stochasticity.

1.2 Motivation and objectives of this thesis

The experiments carried by Hurt et al. (2010) and the mathematical model developed by McCaw et al. (2011) serve as cornerstone for this project. Here, in order to overcome some of the limitations of McCaw et al. (2011), we intend to explain the data with an alternative model with more degrees of freedom, but that still preserves the biological meaning of an influenza transmission, as well as the possibility to capture to very different ecological scenarios of competition between strains.

In this thesis, we develop a simple within-host mathematical model and statistical framework that allows us to estimate the fitness transmission of two strains. Our mechanistic dynamical system approach will advance existing approaches to modelling of mixtures, by including explicit processes on within-host competition between two viral strains. In Chapter 2, we introduce and

analyse the model. Chapter 3 focuses on how we can apply the model to transmission data, and we test its predictions using simulations. In Chapter 4 we develop the data fitting framework that allow us to make inferences and apply simulation approaches to deal with parameter uncertainty. Finally, in Chapter 5 we discuss the implications and limitations of our approach.

Wolfram Mathematica (version 11.3) is used for analytical study of the models, while R (version 3.5.1) is used for numerical analysis. Ordinary differential equations are solved in time using the R function `ode` (package `deSolve`).

This work serves as a general proof-of-concept focusing on asymmetric strain interactions, using programming and statistical tools. It will provide greater biological insight on within-host viral dynamics and propose a new statistical methodology for transmission fitness quantification between viral strains, as well as possible applications to similar dynamic systems that involve competition between species.

Chapter 2

Model framework

A well-known description of competition dynamics between species has been the use of the Lotka-Volterra equations (Murray, 2002). These nonlinear equations are used to describe a dynamical system - a system that changes with time. Models based on these equations had a profound impact on the field of population biology. In this chapter, we describe our model based on competitive Lotka–Volterra equations.

2.1 Lotka-Volterra competition model

A continuous model of ordinary differential equations (ODE) is applied. The differential equations give the rate of change of the two viral populations in study as a function of themselves and of time. This model is deterministic, meaning that for given initial conditions, and fixed parameters, the state of the system will be uniquely determined.

Let $n_1(t)$ and $n_2(t)$ be the number of virions, in a given host, of the mutant strain (strain 1) and the wild-type strain (strain 2), respectively, at time t. They change with time according to the following equations:

$$\frac{dn_1}{dt} = r_1 n_1 - c_{11} n_1^2 - c_{12} n_1 n_2 \tag{2.1}$$

$$\frac{dn_2}{dt} = r_2 n_2 - c_{22} n_2^2 - c_{21} n_1 n_2 \tag{2.2}$$

At any given time, strain i grows at a constant rate r_i, competes with its virions from the same strain with strength c_{ii} and competes with the other strain with strength c_{ij}, for $i = 1, 2$ and $j = 2, 1$. In other words, r_i could be seen as how quickly a strain reaches its carrying capacity, which is here represented as $K_i = \frac{r_i}{c_{ii}}$. To sum up: the first term $(r_i n_i)$ represents

the exponential growth of strain i; the second term ($c_{ii}n_1^2$) corresponds to the density-dependent limitation (logistic growth); and the third term corresponds to the inter-strain competition. All parameter values are assumed to be positive.

2.1.1 Asymptotic analysis of the Lotka-Volterra model

We conduct an asymptotic analysis with the software Mathematica to draw a deeper grasp of the model. To determine what are the end results of this dynamical system, its equilibrium points must be obtained. A system reaches an equilibrium when all the variables that describe its behaviour do not change with time. This is calculated by finding where the differential equations are equal to zero.

The steady-states of this system are given by solving

$$\frac{dn_1}{dt} = 0 \text{ and } \frac{dn_2}{dt} = 0.$$

The fixed points, i.e. solutions of the system, for $t \to \infty$, are then

$$S_1 = \begin{bmatrix} n_1^* = 0 \\ n_2^* = 0 \end{bmatrix}, \quad S_2 = \begin{bmatrix} n_1^* = \frac{r_1}{c_{11}} \\ n_2^* = 0 \end{bmatrix}, \quad S_3 = \begin{bmatrix} n_1^* = 0 \\ n_2^* = \frac{r_2}{c_{22}} \end{bmatrix}, \quad S_4 = \begin{bmatrix} n_1^* = \frac{c_{22}r_1 - c_{12}r_2}{c_{11}c_{22} - c_{12}c_{21}} \\ n_2^* = \frac{c_{21}r_1 + c_{11}r_2}{c_{12}c_{21} - c_{11}c_{22}} \end{bmatrix}$$

For any given values of the parameters, the exact values of n_1^* and n_2^* change, yet these four analytic solutions are maintained. The first solution (S_1) corresponds to a situation where neither the mutant strain nor the wild-type strain exist. This is trivial equilibrium and will be ignored, since it is non-informative and of no interest to our study. In the second case (S_2) the wild-type population is not present while the mutant population persists, while in the third case (S_3) only the wild-type strain persists. In the final fixed point (S_4), the mutant and wild-type populations coexist.

The explicit solutions $n_1(t)$ and $n_2(t)$ are called the trajectories of the system, and different initial conditions can produce different trajectories, but leading always to one of the four equilibria for $t \to \infty$. A key feature in asymptotic analysis is the exploration of the stability in the equilibrium points.

An equilibrium point can be stable or unstable based on the local behaviour of the solutions around the equilibrium point. By perturbing the equilibria, if the trajectories remain near the fixed point, then it is considered stable, and if any of these trajectories do not remain in a neighborhood of the fixed point, it is considered unstable.

The stability of the system's fixed points is formally evaluated using the Jacobian matrix. The Jacobian matrix is a matrix of the partial derivatives of the ODEs with respect to state variables. For our system of ODEs, the Jacobian matrix J is given by

$$J = \begin{pmatrix} \frac{\partial f(n_1,n_2)}{\partial n_1} & \frac{\partial f(n_1,n_2)}{\partial n_2} \\ \frac{\partial g(n_1,n_2)}{\partial n_1} & \frac{\partial g(n_1,n_2)}{\partial n_2} \end{pmatrix}$$

where $f(n_1, n_2)$ and $g(n_1, n_2)$ are the ODEs 2.1 and 2.2, respectively.

To determine the local stability at these points, the Jacobian matrix is evaluated at the equilibria n_1^* and n_2^* at each of S_1-S_4, and then obtain

$$J\big|_{n_1=n_1^*, n_2=n_2^*} = \begin{pmatrix} r_1 - 2c_{11}n_1^* - c_{12}n_2^* & -c_{12}n_1^* \\ -c_{21}n_2^* & r_2 - 2c_{22}n_2^* - c_{21}n_1^* \end{pmatrix}.$$

A equilibrium is stable if every eigenvalue, λ_i, of $J(n_1^*, n_2^*)$ has a negative real part. That is

$$Re\,(\lambda_i) < 0 \quad \forall\, i\,.$$

For the solutions of the system presented above (S_1-S_4), we check the values of the eigenvalues of the Jacobian matrix. Assuming that all parameter values must be positive, we arrive at the following parameter conditions that lead to four different ecological scenarios:

1. If $r_2 < \frac{c_{21}r_1}{c_{11}}$, the solution S_2 is stable, and strain 1 leads to the competitive exclusion of strain 2.

2. If $r_1 < \frac{c_{12}r_2}{c_{22}}$, the solution S_3 is stable, and strain 2 leads to the competitive exclusion of strain 1.

3. If $c_{11}c_{22} > c_{12}c_{21}$ and $\frac{c_{22}r_1}{c_{12}} > r_2 > \frac{c_{21}r_1}{c_{11}}$, the solution S_4 is stable, and strains 1 and 2 co-exist stably.

4. If $c_{11}c_{22} < c_{12}c_{21}$ and $\frac{c_{22}r_1}{c_{12}} < r_2 < \frac{c_{21}r_1}{c_{11}}$ and $r_1 < \frac{c_{12}r_2}{c_{22}}$, the solution S_4 is unstable. This means that, depending on the initial conditions, the system either collapses to the solutions S_2 or S_3 (bistability of the competitive exclusion equilibria).

The parameter conditions of the third scenario can be interpreted as stable coexistence is possible only if the intra-strain competition is stronger than the inter-strain competition. In the fourth scenario, if we start at some particular proportion of the mixture either the mutant (strain 1) or the wild-type (strain 2) may win, eventually leading to exclusion of one strain or of the other strain.

To illustrate the four scenarios of the Lotka-Volterra model with two strains within a given host (Figure 2.1), we chose the parameter values presented in Table 2.1.

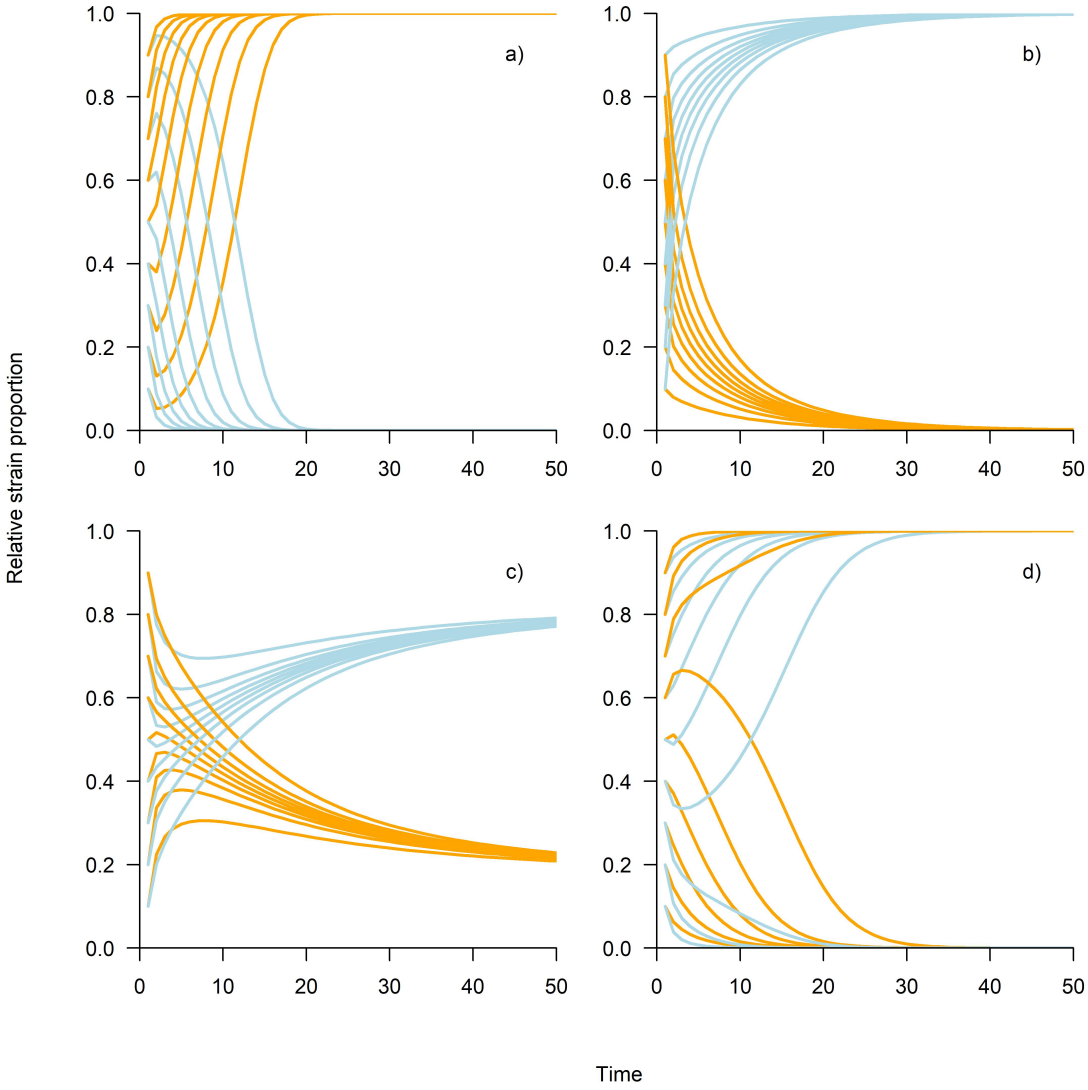

Figure 2.1: Illustration of the dynamics over time of the deterministic model scenarios derived from asymptotic analysis: a) mutant (strain 1) out-competing wild-type (strain 2), b) strain 2 out-competing strain 1, c) coexistence, d) bistability. Each line represents a different starting condition (i.e. different starting proportion of the mutant, $n_1(t)/(n_1(t)+n_2(t))$) for the mutant strain (orange lines) and the wild-type strain (blue line). The values of the parameters used are shown in Table 2.1.

2.2 Rescaled model

In order to simplify the model and ease the later parameter estimation process, we reduced the number of parameters by doing a nondimensionalization. It greatly simplifies the model by re-writing the parameters as unitless combinations, thus reducing the total number of parameters to estimate.

Table 2.1: Model parameter values for infection dynamics simulation with time within a host.

Parameter	Scenario 1 (strain 1 wins)	Scenario 2 (strain 2 wins)	Scenario 3 (coexistence)	Scenario 4 (bistability)
r_1	0.7	0.5	0.1	0.7
r_2	0.2	0.6	0.7	0.4
c_{11}	0.4	0.7	0.6	0.4
c_{12}	1.0	0.8	0.2	0.2
c_{21}	0.6	0.2	0.3	0.6
c_{22}	0.6	0.8	0.6	0.1

We define the new variables

$$u_1 = \frac{n_1}{r_1/c_{11}} \text{ and } u_2 = \frac{n_2}{r_2/c_{22}} \tag{2.3}$$

and rescale the parameters as:

$$\rho = \frac{r_2}{r_1}, \;\; a_{12} = c_{12}\frac{r_2/c_{22}}{r_1/c_{11}}, \;\; a_{21} = c_{21}\frac{r_1/c_{11}}{r_2/c_{22}}. \tag{2.4}$$

The rescaled model is then ruled by the following equations:

$$\frac{du_1}{d\tau} = u_1(1 - u_1 - a_{12}u_2) \tag{2.5}$$

$$\frac{du_2}{d\tau} = \rho u_2(1 - u_2 - a_{21}u_1) \tag{2.6}$$

Notice, the time in this rescaled model is also scaled to the new time-scale $\tau = r_1 t$. So, effectively, we drop from 6 parameters to 3: ρ is the ratio of the growth rates of the two strains, and a_{12} and a_{21} are the relative competition indices.

2.2.1 Asymptotic analysis of the rescaled model

The ecological scenarios of this system are then determined only by the values of the parameters a_{12} and a_{21}, as summarized in Figure 2.2. The phase plots of the corresponding scenarios are shown in Figure 2.3, now only dependent on the relative magnitudes of two parameters, a_{12} and a_{21}.

The steady-states of this system are given by:

$$\frac{du_1}{d\tau} = 0 \text{ and } \frac{du_2}{d\tau} = 0.$$

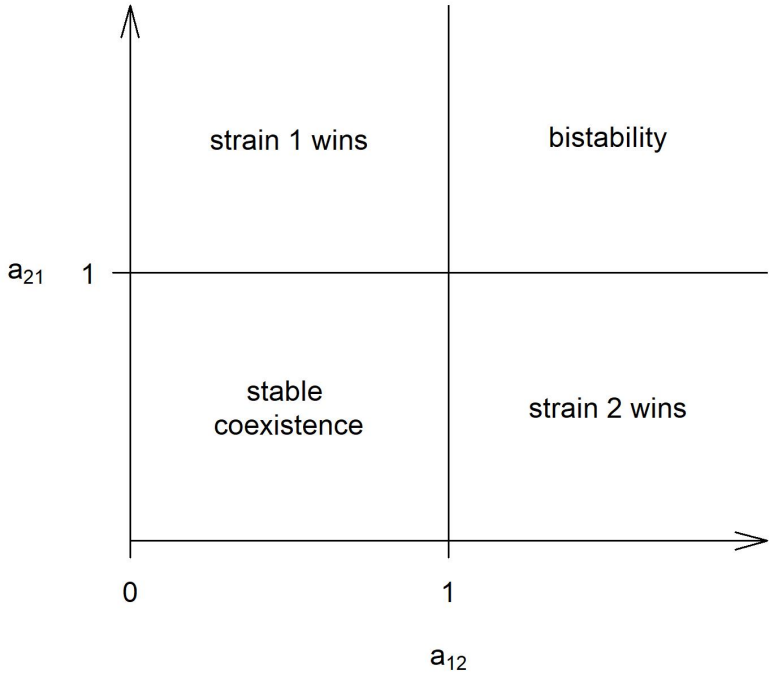

Figure 2.2: Summary of the general parameter conditions for the system's scenarios. The four characteristic scenarios in the rescaled model are given exclusively by the values of a_{12} and a_{21}.

The solutions to the system are then

$$S_1 = \begin{bmatrix} u_1^* = 0 \\ u_2^* = 0 \end{bmatrix}, \quad S_2 = \begin{bmatrix} u_1^* = 1 \\ u_2^* = 0 \end{bmatrix}, \quad S_3 = \begin{bmatrix} u_1^* = 0 \\ u_2^* = 1 \end{bmatrix}, \quad S_4 = \begin{bmatrix} u_1^* = \frac{1-a_{12}}{1-a_{12}a_{21}} \\ u_2^* = \frac{1-a_{21}}{1-a_{12}a_{21}} \end{bmatrix}$$

These four solutions are qualitatively equivalent to those of the full Lotka-Volterra model, yet much simpler in terms of parameters. Again, to determine the local stability at these points, the Jacobian matrix is calculated and evaluated at the equilibria u_1^* and u_2^*.

$$J\big|_{u_1=u_1^*, u_2=u_2^*} = \begin{pmatrix} 1 - 2u_1^* - a_{12}u_2^* & -a_{12}u_1^* \\ -a_{21}\rho u_2^* & \rho(1 - a_{21}u_1^* - u_2^*) - \rho u_2^* \end{pmatrix}$$

An asymptotic analysis carried in Mathematica allows the derivation of the system's parameter conditions, leading to the same four characteristic scenarios:

1. If $a_{12} < 1$ and $a_{21} > 1$, the solution S_2 is stable, and strain 1 leads to the competitive exclusion of strain 2.

2. If $a_{12} > 1$ and $a_{21} < 1$, the solution S_3 is stable, and strain 2 leads to the competitive exclusion of strain 1.

3. If $a_{12} < 1$ and $a_{21} < 1$, the solution S_4 is stable, and both strains coexist stably.

4. If $a_{12} > 1$ and $a_{21} > 1$, the solution S_4 is unstable, and S_2 or $S3$ are the stable equilibria the system tends to, depending on the initial conditions.

Next we apply the phase plane method to illustrate the solutions in a 2-D space over time. Equations 2.5 and 2.6 can be written using vector notation as

$$\dot{\boldsymbol{u}} = \boldsymbol{h}(\boldsymbol{u})$$

where $\boldsymbol{u} = (u_1, u_2)$ represents a point in the phase plane, and $\boldsymbol{h}(\boldsymbol{u}) = (f(u_1), f(u_2))$. $\dot{\boldsymbol{u}}$ is the velocity vector at that point. Different initial conditions correspond to different points and different trajectories. Using the phase plane as a visualisation tool is intuitive to get an understanding of the final outcomes of the system for each of the four ecological scenarios.

In Figure 2.3 the trajectories of the rescaled system are portrayed in the phase plane, using the parameter values of Table 2.2. Over time, the solutions tend to the equilibria S_2-S_4.

Table 2.2: Model parameter values for transmissions dynamics simulation from a donor to recipient.

Parameter	Scenario 1 (strain 1 wins)	Scenario 2 (strain 2 wins)	Scenario 3 (coexistence)	Scenario 4 (bistability)
ρ	1	1	1	1
a_{12}	0.5	1.5	0.5	1.5
a_{21}	1.5	0.5	0.5	1.5

2.3 Conclusions

Now we have a model that describes in a more complex way the dynamics between two strains of viruses. Our approach to the transmission of competitive-mixtures interprets competition as the main ecological force driving the differential transmission fitness. The increase in model parameters and complexity pays off with a bigger flexibility. From our analytical and numerical explorations, we presented two scenarios of competitive exclusion, one of coexistence and one of bistability. These last two could not be captured with the model presented by McCaw et al. (McCaw et al., 2011). In the following chapter, we will describe how we can apply this framework to the data structure provided by Hurt et al. (Hurt et al., 2010) and how we validate the findings through simulations.

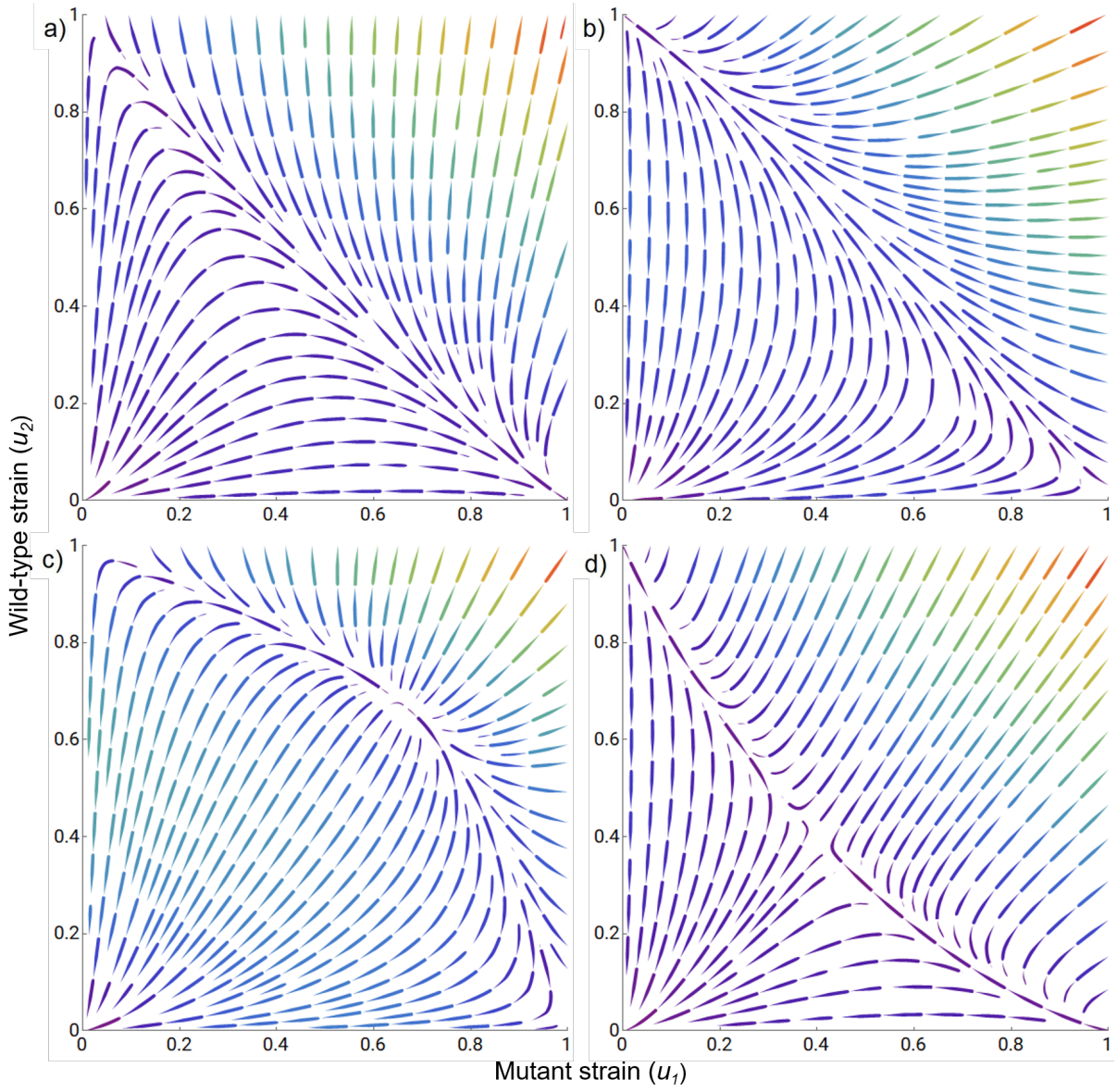

Figure 2.3: Phase plane plots of the model scenarios: a) strain 1 outcompetes strain 2, b) strain 2 outcompetes strain 1, c) coexistence and d) bistability of the competitive exclusion equilibria. The values of the parameters used are shown in Table 2.2.

Chapter 3

Model evaluation by simulation studies

In this thesis, we lean on the transmission of influenza strain mixtures, focusing on two strains that differ in their resistance or susceptibility to an antiviral. By assigning the number of virions at time $t = 0$ as being in the donor and at time $t = 1$ being in the recipient we are able to interpret the transmission dynamics from our model. The time has been rescaled in units of observation. Given the full Lotka-Volterra model we determine:

$$p_d = \frac{n_1(0)}{n_1(0) + n_2(0)} \quad \text{and} \quad p_r = \frac{n_1(t)}{n_1(t) + n_2(t)}$$

where p_d and p_r are the proportion of the mutant strain (strain 1) in the donor and in the recipient.

Writing these proportions in terms of the rescaled model we obtain:

$$p_d = \frac{u_1(0)}{u_1(0) + u_2(0)\beta} \quad \text{and} \quad p_r = \frac{u_1(\tau)}{u_1(\tau) + u_2(\tau)\beta}$$

where we assume that $\tau = 1$, corresponding to a snapshot observation in the recipient, respectively, and $\beta = K_1/K_2$ is the ratio of the carrying capacities.

Figure 3.1 illustrates the expected transmission behaviours for the asymptotic scenarios described previously, using the parameter values of Table 3.1 for the rescaled model.

Again, four characteristic scenarios emerge, even when a single time point is evaluated in the recipient, as a function of different initial mixture proportions originating from the donor. Curves above the diagonal of the plot mean the mutant strain out-competes the wild-type, during transmission from donor to recipient, while curves below the diagonal mean the opposite. Curves cross the diagonal horizontally in the scenario of stable coexistence between the two strains, where the crossing occurs around the expected stable coexistence. Finally, curves cross

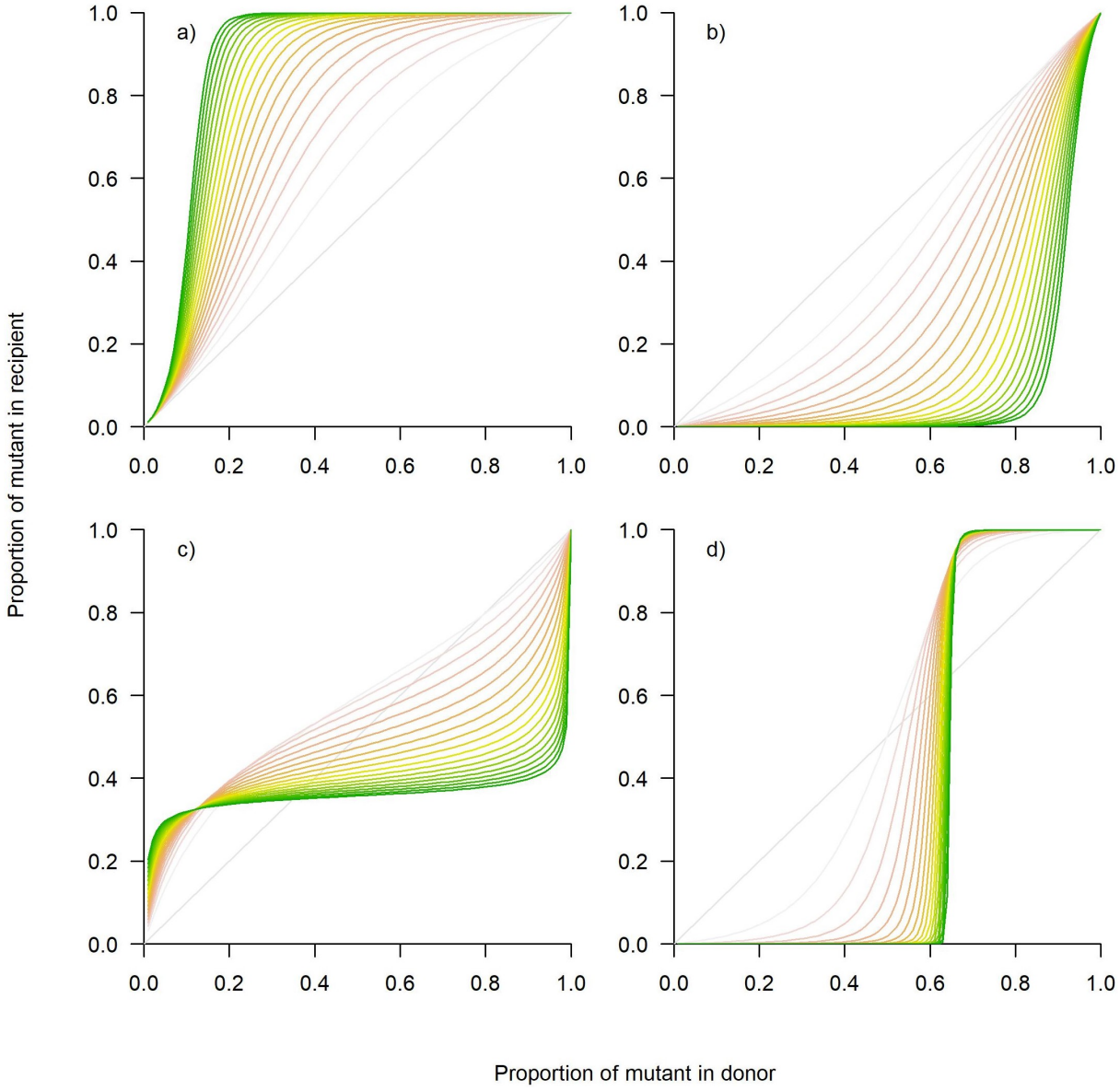

Figure 3.1: Transmission illustrations of the model scenarios: a) strain 1 out-competing strain 2, b) strain 2 out-competing strain 1, c) coexistence, d) bistability. Each line represents a different timepoint solution of the ODEs, and the greener it is, the closest it is to the equilibrium. The values of the parameters used are presented in Table 3.1.

Table 3.1: Rescaled model parameter values for simulations of transmissions from a donor to recipient.

Parameter	Scenario 1 (strain 1 wins)	Scenario 2 (strain 2 wins)	Scenario 3 (coexistence)	Scenario 4 (bistability)
ρ	0.9	1.8	1.5	1.0
a_{12}	0.4	1.4	0.5	1.7
a_{21}	1.6	0.7	0.4	1.6
β	1.0	1.0	1.0	1.0

the diagonal vertically in the scenario of bistability, where the crossing occurs at the separatrix (that separates the results of the unstable equilibrium) for those parameters.

Now that we can apply the model to datasets like those given by McCaw et al. (2011), we describe briefly the optimization algorithm used for the model fitting to data.

3.1 Optimization algorithm

We use non-linear optimization, which intends to find the combination of parameters that gives the best solution to a given data set. In our case, we want to find the minimum between our model predictions and the data. The proportions in the donor are considered given (fixed) and used as initial conditions for the dynamical system. The proportions in the recipient are matched with model-predictions evaluated at time $\tau = 1$.

The objective of our optimization algorithm is to minimize the mean squared errors (MSE), given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(p_r(i) - \hat{p}_r(i))^2$$

where $p_r(i)$ represent each of the observed values of p_r, the proportion of strain 1 in the recipient, and $\hat{p}_r(i)$ are the predicted values for p_r from the deterministic Lotka-Volterra model.

The MSE minimization is done using the R optimization function `optim`. The `optim` function minimises the cost function by varying its parameters, starting at the provided initial values. With some functions, particularly functions with many minimums, the initial values have a great impact on the converged point. Other optimization functions were tested and `optim` produced the best results while being the easiest to use (see Supplemental Section S1). The `optim` function can accept many different types of methods of optimization, from simple quasi-Newton methods to more complex procedures. Unless stated otherwise, all optimizations were conducted with the "L-BFGS-B" method (Byrd et al., 1995), since it accepts box constraints, that is, each variable can be given a lower and/or upper bound. This is relevant since for all models applied the parameters must have positive values.

As mentioned before, in the optimization function the initial guess for the parameters has an effect on the optimization procedure. This is illustrated in Figure 3.2. This is due to the fact that the function to be minimized has many local optima, and the optimization does not converge to the global minimum.

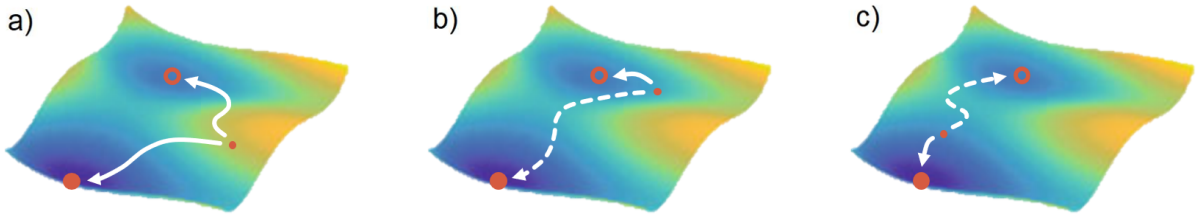

Figure 3.2: Illustration of local and global minimum in an optimization procedure. The small red dot represents the starting guess for the parameters, the filled red circle represents the global minimum of the error space and the red circumference is a local minimum. The optimization changes iteratively the values of the starting guesses until it converges to a minimum. In a) the optimization can converge from that starting point to either a local or a global optimum. In b) a convergence to a local minimum is more likely (or at least much faster) than a convergence to the global minimum. In c) the starting guess for the parameters favors a fast convergence to the global minimum. Note that this is a mere abstraction: the error space of our system cannot be visualised due to the multi-dimensionality created by having more than two parameters. Adapted from (Fröhlich et al., 2019).

3.2 Simulation framework

We create a framework to simulate dynamics from the mathematical model (equations 2.5 and 2.6) and try to estimate the parameters that generated these simulations. The idea is to test whether this procedure can recover/identify the 'true' strain competition and growth parameters from observations and just one time point ($\tau = 1$) but starting from many initial conditions (mixture proportions). This section describes the methodology applied.

Table 3.2: Parameter intervals used for the simulation algorithm. Each interval for random parameter generation correspond to the intended scenario.

Scenario	ρ	a_{12}	a_{21}	β
1 (*strain 1 excludes 2*)	[0.01, 2]	[0.01, 1]	[1, 2]	[0.01, 2]
2 (*strain 2 excludes 1*)	[0.01, 2]	[1, 2]	[0.01, 1]	[0.01, 2]
3 (*coexistence*)	[0.01, 2]	[0.01, 1]	[0.01, 1]	[0.01, 2]
4 (*bistability*)	[0.01, 2]	[1, 2]	[1, 2]	[0.01, 2]

The values of the simulated parameters, $\boldsymbol{\theta} = (\rho, a_{12}, a_{21}, \beta)$, shown in Table 3.2, were used to obtain the typical scenario curves at $\tau = 1$. Our model fits to the data from the experiments of Hurt et al. (Hurt et al., 2010) using the `optim` function (with method L-BFGS-B and no constrains for the parameter values). In order to avoid the `optim` function getting stuck on bad initial guesses, the initial $\boldsymbol{\theta}$ guesses were fixed with the values presented in Table 3.3. We run 100 random iterations for each scenario separately, in order to study if there is any particular estimation bias, i.e. if there is any reason to assume a given scenario could have a better/worse quality of fit.

Table 3.3: Initial guesses for the parameters used for the simulation algorithm. These values were chosen so that the optimization algorithm has a starting guess already within the intended scenario.

Scenario	ρ	a_{12}	a_{21}	β
1	1	0.5	1.5	1
2	1	1.5	0.5	1
3	1	0.5	0.5	1
4	1	1.5	1.5	1

3.2.1 Measures of quality-of-fit

The quality of the fits will be measured by the mean squared errors (MSE) and by Δ – how far the estimated parameters are from the simulated ones – given by:

$$\Delta_i^m = \frac{\widehat{\mathbf{\Theta}}_i^m - \mathbf{\Theta}_i^m}{\mathbf{\Theta}_i^m}$$

where $\mathbf{\Theta}$ is the matrix of simulated parameters generated randomly within the intervals specified in Table 3.2, i refers to the parameter of the model, m refers to the number of the iteration, $\widehat{\mathbf{\Theta}}$ is the matrix of best parameter estimates given by the model fit:

$$\mathbf{\Theta} = \begin{bmatrix} \rho^1 & a_{12}^1 & a_{21}^1 & \beta^1 \\ \rho^2 & a_{12}^2 & a_{21}^2 & \beta^2 \\ \vdots & \vdots & \vdots & \vdots \\ \rho^m & a_{12}^m & a_{21}^m & \beta^m \end{bmatrix}, \qquad \widehat{\mathbf{\Theta}} = \begin{bmatrix} \widehat{\rho}^1 & \widehat{a}_{12}^1 & \widehat{a}_{21}^1 & \widehat{\beta}^1 \\ \widehat{\rho}^2 & \widehat{a}_{12}^2 & \widehat{a}_{21}^2 & \widehat{\beta}^2 \\ \vdots & \vdots & \vdots & \vdots \\ \widehat{\rho}^m & \widehat{a}_{12}^m & \widehat{a}_{21}^m & \widehat{\beta}^m \end{bmatrix}$$

The precision of the model is measured as the spread of the variance of Δ^m, the mean over i (for each iteration, $\Delta^m = 1/4 \sum_{i=1}^m \Delta_i^m$), while the accuracy is the bias of the mean (i.e. how close is the mean to 0).

The relation between the MSE and Δ over all iterations of the model fit can also give us some insights on the behaviour of the model and estimation procedure. For example, if the model has very low values of MSE but high values of Δ, it means that the model can produce curves that are very close to the data even if the parameters are very different than those which generated it. In other words, this relation can warn us of parameter identifiability issues (see Supplemental Section S2 for details on local and global optima assessment).

3.3 Simulation results

To test the reliability of the model predictions we conducted a simulation approach to validate the model. We now present the results from the model validation. Figure 3.3 shows the distributions of the mean squared errors (MSE) - the error between the simulated data points and the fitted ones - and of the parameter accuracy (Δ) - the difference between the simulated parameters and the estimated parameters.

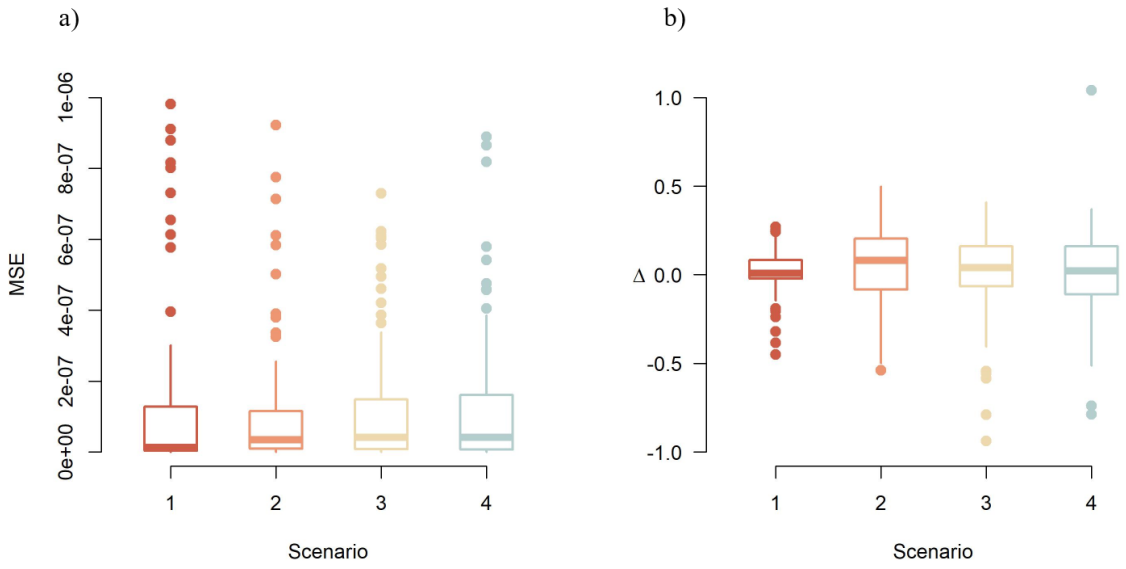

Figure 3.3: Measures of quality of fit and parameter accuracy results of the model validation procedure in the absence of sampling error. a) Boxplot of MSE across the four scenarios from simulations (100 iterations per scenario). The boxplots show the median, lower and upper quartiles, and the extreme lines reaching the minimum and maximum values excluding the outliers (shown as dots). To facilitate visualization, some outliers are not shown (5 from scenario 1, 1 from scenario 2, 6 from scenario 3, 7 from scenario 4). b) Boxplot of Δ^m across the four scenarios. To facilitate visualization, some outliers are not shown (1 from scenario 1 and 1 from scenario 4).

Both measures have a mean close to zero, indicating that the model reaches a good quality of fit as well as proximity to the real parameters.

3.4 Conclusions

We now have a mechanistic model that can be applied to the data structure of competitive mixtures between strains. Using optimization routines to estimate the parameters, we will be able to, in this way, infer parameters of the within-host dynamics that could not be captured with a simpler exponential model, like that of McCaw et al. (2011). Additionally, by simulating data with known parameters and fitting the model to this artificial data, we could quantify the quality of fit as well as the parameter accuracy for all the scenarios of competitive outcomes of

the two strains. Both measures of difference between artificial and model predictions were close to zero, indicating a convergence to the real transmission dynamics. This means we can be fairly confident that the estimates provided by our model fitting to data should be close enough to the real-world dynamics.

Chapter 4

Model fitting to data and uncertainty quantification

This chapter concerns the key stage of the mathematical modelling procedure: the model fitting to real data and subsequent interpretation of the parameter estimates. Additionally, we describe the different approaches to tackle the uncertainty in parameter estimation for this data.

As a point of reference, we first replicate the findings of McCaw et al. (2011) by applying their model (equation 1.1) to the data (Table 4.1). Figure 4.1 shows the best model fit (red line) and the corresponding 95% confidence intervals for the relative fitness differences coefficient s. The best estimate is obtained through standard optimization procedure using `optim` function in R, as described in Chapter 3, and as will be applied later to our model as well. The resulting best estimate is $s = -0.2598$ $(-1.3509, 0.8329)$[1], and the MSE of this model fitting is 0.037. The estimate of s was consistent with the one presented by the authors, and indicates a slight growth advantage to the mutant, but with very large uncertainty, including a converse scenario. We will compare this results with our model estimates at the end of this chapter. Given that our model has more parameters, and hence a greater flexibility, we expect to achieve a better quality of fit.

Table 4.1: Data used for model fitting. The values correspond to measured proportions of the mutant strain in the donor (p_d) and recipient (p_r) ferrets in the transmission events carried by Hurt et al. (Hurt et al., 2010).

	Observed proportions of the mutant strain								
p_d	0.00	0.08	0.09	0.12	0.35	0.60	0.82	0.95	0.99
p_r	0.00	0.00	0.20	0.43	0.68	0.33	0.63	0.99	0.94

[1]These confidence intervals correspond to the ones presented by the authors, since their bootstrap approach was not replicated.

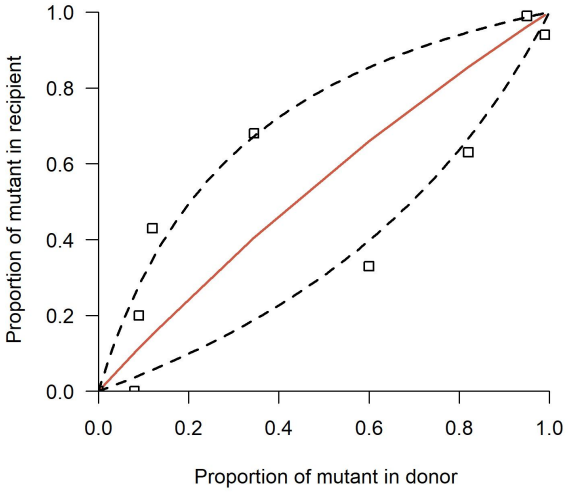

Figure 4.1: Replication of the best model fit (from McCaw et al. (McCaw et al., 2011)) to H274Y data (squares) is shown by the red solid line using the function shown previously. Best estimate for the parameter s obtained through standard optimization procedure using `optim` function in R ($s = -0.2598$), with MSE = 0.037. 95% confidence intervals for s, $(-1.3509, 0.8329)$, are shown by the dashed lines.

4.1 Model fitting to real data

Next we fitted our model, in its rescaled version (equations 2.5 and 2.6), to the H274Y experimental data of McCaw et al. (2011) (Table 4.1). The model fit is shown in Figure 4.2. We can already see that the model fits better to the data. This is confirmed by comparing the MSE of this fit ($\widehat{MSE} = 0.0129$) to the MSE of the McCaw et al. (2011) (MSE = 0.037). The more complex model yields a 3 times lower error, thus 3 times better quality of fit than the simpler exponential growth formulation. The best-fitting parameters, $\hat{\boldsymbol{\theta}} = (\hat{\rho}, \hat{a}_{12}, \hat{a}_{21}, \hat{\beta})$, are shown in Table 4.2, as well as a summary of the fitting procedure constrains on the parameters and initial guesses for the parameters. For R code see Supplemental Section S4.

Table 4.2: Initial parameter guesses $\boldsymbol{\theta}_0$, parameter constraints used in the optimization algorithm, and model parameter estimates ($\hat{\boldsymbol{\theta}}$) from the data fitting to the H274Y data.

$\boldsymbol{\theta}$	$\boldsymbol{\theta}_0$	$\boldsymbol{\theta}$ bounds	$\hat{\boldsymbol{\theta}}$
ρ	1	$[0, \infty[$	96.325
a_{12}	1	$[0, \infty[$	0.759
a_{21}	1	$[0, \infty[$	0.951
β	1	$[0, \infty[$	0.999

The rescaled Lotka-Volterra model fit is consistent with the scenario of coexistence within-

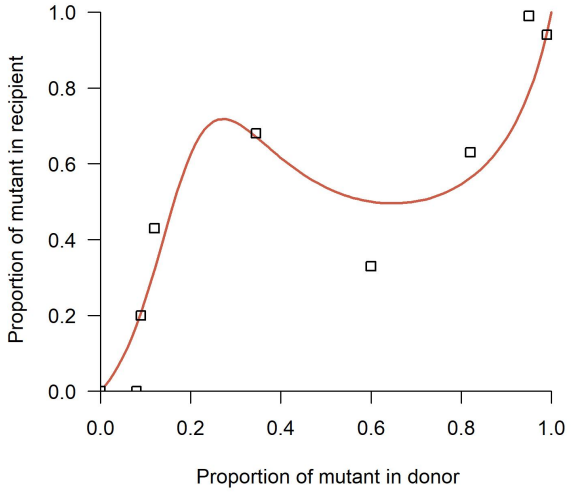

Figure 4.2: Best model fit to H274Y data (squares) is shown by the red solid line, as estimated by the rescaled model equations 2.5 and 2.6, with MSE = 0.0129. The best estimates for the parameters, obtained through optimization, are shown in Table 4.2.

host between the two strains of influenza, since both a_{12} and a_{21} are below 1 (see parameter conditions in Section 2.2.1). We can see that the fitness advantage of the wild-type (strain 2) in growth ($\hat{\rho} \approx 95$) is counteracted by a competition advantage of the mutant (strain 1) in inter-strain interactions ($\hat{a}_{12} < \hat{a}_{21}$). The model also infers that the ratio of within-host carrying capacities of the two strains is around 1 ($\hat{\beta} = 1$), thus suggesting that the resource limitation for growth when alone, probably acts similarly on each strain. In fact, these estimates capture the pattern of initial mutant growth advantage when rare, because it experiences less competition from the wild-type (points above the diagonal for low mixture proportions), and initial growth disadvantage when frequent (points below the diagonal for higher mixture proportions). Coexistence then results from inter-strain competition being relatively weaker than intra-strain competition. The net effect of these parameter estimates is that the two strains are expected to coexist within host at equilibirum, an outcome that could also be seen as very little fitness difference between the two strains (e.g. interpretation of results by McCaw et al. (2011)).

Thus we can see that this model is able to flexibly capture the diagonal crossing of data points at a certain mutant proportion. Additionally, the estimated parameters predict that over time the competition between strains should settle at

$$u_1^* = \frac{1 - a_{12}}{1 - a_{12}a_{21}}, u_2^* = 1 - a_{12}u_1^*.$$

which leads to an equilibrium mutant proportion of

$$p_r(t) \xrightarrow{t\to\infty} p_r^* = \frac{u_1^*}{u_1^* + u_2^*/\beta} \approx 0.83.$$

Of course, this limit may be never reached exactly, due to the action of host immunity, which will be triggered at some later point during infection and clear the pathogen population. Since we are not modeling the acute phase of infection with our formulation (time $\approx$ 1 day), this partition of viral population between mutant and wild type strain is expected in the intermediate time frame before immunity is activated, typically 3 days after infection (Tamura and Kurata, 2004).

After having obtained parameter estimates with our more complex model, we are interested in quantifying the uncertainty around the parameters. Since the optimization routine that we apply does not provide readily confidence intervals for our best-estimates, we choose to use a bootstrap approach. However, common bootstrap methods that involve simply resampling data or resampling residuals, affect the data structure, and cannot be applied to our data.

Thus, we adopt two strategies to quantify the uncertainty around the parameter estimates. Both rely on generating artificial data based on the first best estimates ($\hat{\boldsymbol{\theta}}$, Table 4.2), to simulate many repeated equivalent experiments. They differ, nonetheless, in the way they create synthetic data: one is based on uncertainty caused by experimental measurements (experimental error), while the other focuses on uncertainty caused by stochasticity during transmission (bottleneck effect). We will describe in detail in the next two sections the procedures applied and the results obtained from these two methods.

4.2 Uncertainty caused by experimental error

The experimental setup of Hurt et al. (Hurt et al., 2010) may be subject to some variation in the measured proportions of the two strains due to the intrinsic stochasticity of the sampling procedure. To reflect this source of uncertainty, we apply this approach of generating artificial data based on the sampling error.

To simulate stochastic sampling error, we generate 50 random points in circles around the original data points within a certain radius r that reflects the maximum error according to:

$$(p'_d - p_d)^2 + (p'_r - p_r)^2 \leq r_{error}^2$$

where (p'_d, p'_r) are the coordinates of the generated points, (p_d, p_r) are the coordinates of the

original transmission events (see Table 4.1), and r_{error} is radius of the circle. A larger radius means a larger sampling error as shown in Figure 4.3, thus causing more variability in the data.

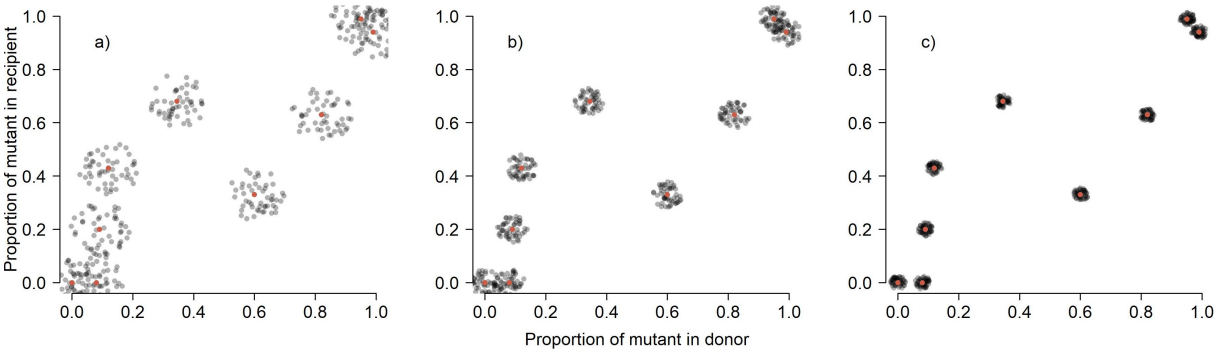

Figure 4.3: Artificial data (black dots) generation based on experimental sampling error around the original data points (red dots). 50 random points are generated from a Uniform distribution using `runifdisc` R function with different radius sizes: a) $r_{error} = 0.1$, b) $r_{error} = 0.05$, c) $r_{error} = 0.025$.

Since it is more reasonable to assume an intermediate sampling error, we fixed a radius of $r_{error} = 0.05$ and generated data uniformly around the original data points. We then fit our model to such sets of artificial data obtained in this way (Figure 4.4 a)). From the set of 50 model fits to the simulated data (gray curves in Figure 4.4 b)), we compute the 95% CI to achieve the empirical confidence region for model predictions (Figure 4.4 c)).

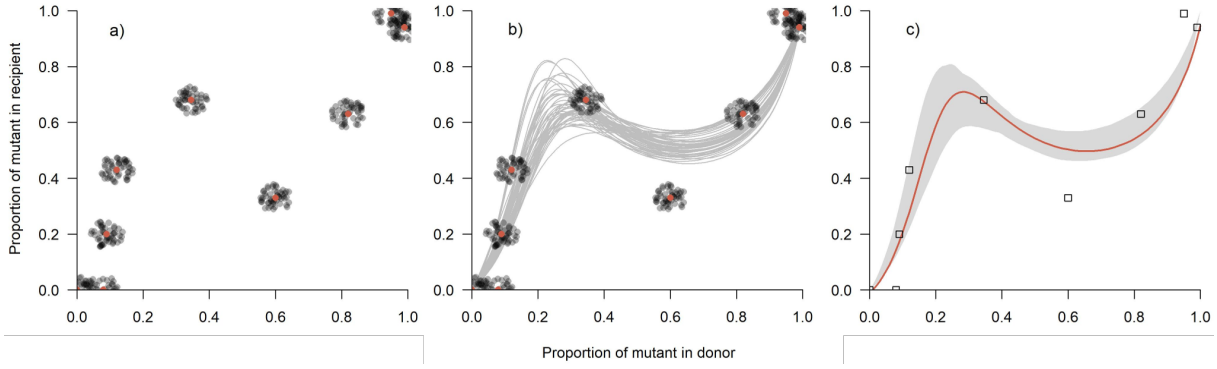

Figure 4.4: Results of fitting the model to simulated data based on sampling error. a) Generated artificial data (50 datasets - black dots) based on intermediate sampling error ($r_{error} = 0.05$) around the original data points (red dots). b) Model fits to simulated data (grey lines). c) 95% empirical confidence region constructed from the model fits (grey region).

In addition, from this procedure we gather the CIs for the parameters, quantifying in this way parameter uncertainty. These are obtained from the 95% empirical quantiles of the best estimates for the parameters from each model fit (see Table 4.3).

Table 4.3: Model parameter estimates from the model fitting to the H274Y data, $\hat{\boldsymbol{\theta}}$, and parameter 95% confidence intervals obtained from the empirical parameter distributions, using the simulation approach based on sampling error of 5% around the original data observations.

$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	95% CI
ρ	96.325	(96.325, 96.326)
a_{12}	0.759	(0.702, 0.857)
a_{21}	0.951	(0.938, 0.971)
β	0.999	(0.996, 1.004)

4.3 Uncertainty caused by the bottleneck effect

In this section, we describe how uncertainty caused by the number of total virions transmitted, N, affects model fitting and parameter estimates. This is another potential source of error, different from the experimental error considered in the previous section.

Thus, we extended the model to include the possible stochastic effects of transmission bottleneck. Modeling explicitly the bottleneck size, i.e. the number of virions transmitted N, implies adding a stochastic sampling step in the x- component of the data. In the model we thus implement the following change for initial conditions, depending on N, based on the binomial model:

$$u_1 = \frac{X}{N}, \qquad u_2 = \frac{1 - u_1}{\beta} \tag{4.1}$$

where $X \sim Binom(N, p_d)$, p_d is the mutant proportion in the donor initiating infection in the recipient, and N is the total viral population size where such proportion is sampled. Different values of N affect the relative proportion of the two strains that is transmitted from one host to another akin to the effect of a transmission bottleneck. Simulations with a different (u_1, u_2) at time $\tau = 0$ lead to different (u_1, u_2) at time τ, even if $\boldsymbol{\theta}$ is kept fixed.

We expect that if there is a distribution of N, it will influence the effective initial conditions in the recipient and consequently affect the final dynamics. Figure 4.5 illustrates schematically the steps applied to simulate the effects of N, using $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, and obtain the confidence region for our model predictions.

We will describe in the following sections how we can use generated data to estimate a plausible distribution for N as well as the 95% CI for $\boldsymbol{\theta}$.

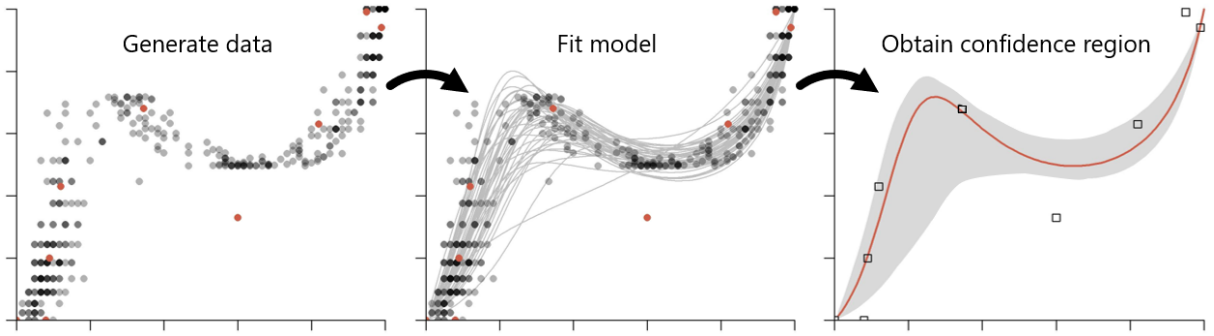

Figure 4.5: Illustration of the steps to simulate the effect of N, apply model fit to simulated data based on N and obtain $\boldsymbol{\theta}$ CIs. The x-component of the artificial data is generated according to condition 4.1 and the y-component by model simulation using for $\boldsymbol{\theta}$ the best estimate, $\hat{\boldsymbol{\theta}}$, obtained in Section 4.1. By selecting appropriate subsets of such artificial data and associated model fits, we can also estimate a distribution for N. The model is then refitted sequentially to the filtered data to obtain the final confidence region for parameters and model predictions.

4.3.1 Filtering artificial data based on distance to the real data

In this section, we will generate data using uncertainty caused by the bottleneck effect, and subsequently filter the data based on the quality-of-fit. We simulated the effect of different N with fixed parameter values, in this case, the best-fitting parameters, $\hat{\boldsymbol{\theta}} = (\hat{\rho} = 96.325, \hat{a}_{12} = 0.759, \hat{a}_{21} = 0.951, \hat{\beta} = 0.999)$. Hence we aim to investigate how much variability this new model element would introduce into the system's dynamics and predictions.

Indeed, simulations with different N show dramatic changes in the predictions for recipient mutant proportions, in particular very high stochasticity for low N and very little stochasticity (close to the deterministic line) for large N (Figure 4.6), as expected from the law of the large numbers.

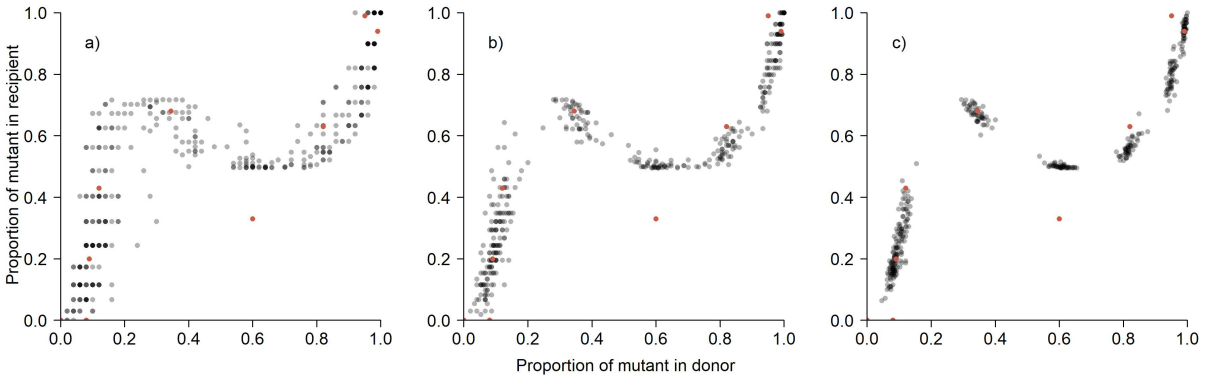

Figure 4.6: Effect of N on simulated data generation. Fixing $\boldsymbol{\theta}$ as the best estimates for the parameters obtained in Section 4.1, we generate data based on the Binomial resampling (condition 4.1) using different values of N: a) $N = 50$, b) $N = 150$, c) $N = 500$.

We simulate the model for $M = 25$ different values of N within the range $[5, 500]$, and using

$\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$. Among these, we filtered those simulations that deviate little from the true data. We define the deviation from the true data, for each run r, as

$$D(N) = \frac{1}{k}\sum_{j=1}^{k}\left(\hat{p}_r(N, \hat{\boldsymbol{\theta}})(j) - p_r(j)\right)^2 \tag{4.2}$$

where $\hat{p}_r(N, \hat{\boldsymbol{\theta}})(j)$ is the vector estimated mutant proportion in the recipient for that given value of N, fixing $\boldsymbol{\theta}$ as $\hat{\boldsymbol{\theta}}$ from Table 4.2, and $j = (1, ..., k)$ refers to each of the 9 transmission events. p_r is the vector of $k = 9$ observed mutant proportion in recipient from the true data (Table 4.1). For each value of N, we conduct $m = 20$ runs to account for stochasticity from N alone.

We then proceed to select those runs that have a D within 20% of the MSE of the original model fitting ($\widehat{MSE} = 0.0129$)

$$\frac{D_i - MSE}{MSE} < 0.2, \qquad i = 1, ..., m. \tag{4.3}$$

We compute the proportion of these runs for each N. We then select the value for N that yields the highest proportion of runs that satisfy condition 4.3. This process is repeated 100 times, leading to a set of selected 'optimal' values of N. This is illustrated in Figure 4.7. In other words, for the values of N within $[5, 500]$, we keep those that lead on average to a higher proportion of iterations with D within 20% of the MSE of our original model fit.

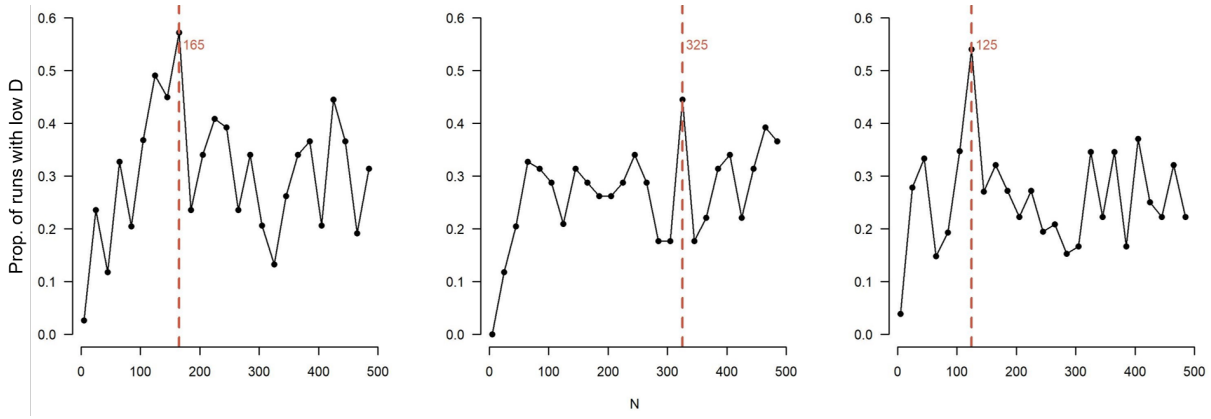

Figure 4.7: Preliminary filtering of N based on the proportion of simulation runs with low D, i.e. that satisfy condition 4.3. On each filtering iteration, the optimal value for N that maximizes the proportion of runs with low D. The collection of all these 'best' values of N can be used as an estimated distribution of N.

This filtering criterion produced a distribution for N skewed towards high values (Figure 4.8). Unsurprisingly, if the selection criterion of $\hat{\boldsymbol{\theta}}$-generated data is based on the proximity to the true data, the law of large numbers guarantees that a higher value for N should be closer to the best deterministic line (Figure 4.1).

Such bias towards a higher N in the inferred distribution leads to a thinner confidence region around the best estimate for the transmission curve and, consequently, less data points

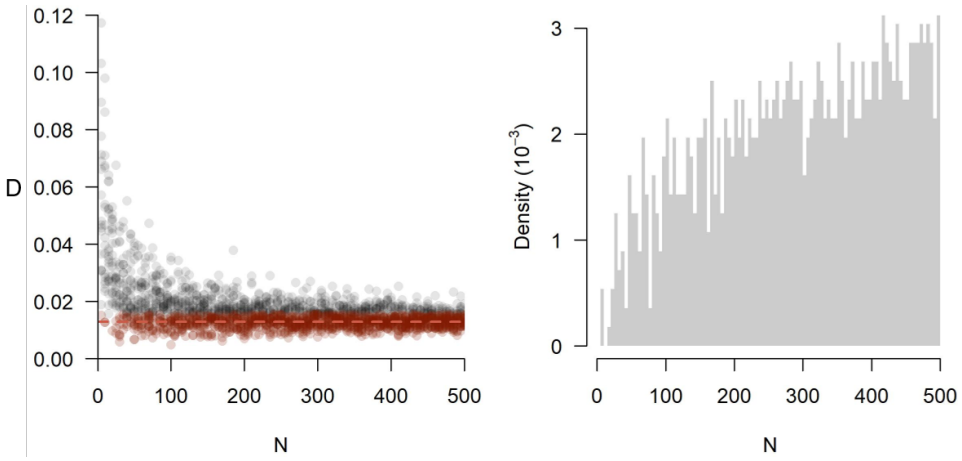

Figure 4.8: Simulations with different N produce different D. a) Scatter plot of D from 20 iterations for each value of N in the range $[5, 500]$. The red dotted line is the MSE of the original data fitting. b) Density of N values that are accumulated from sequential iterations if they satisfy the condition 4.3 (red dots in a)).

are included by N-driven stochasticity. To counteract this limitation that concerns data coverage, we will consider an alternative criterion to filter the N-generated data to accommodate more the variability in the data structure.

4.3.2 Filtering artificial data based on data coverage

To allow for bottleneck-induced stochasticity improve model-fit to data, we consider a second criterion, based on how many of the original data points are contained within the 95% stochastic realizations for a given N.

In Figure 4.9 we illustrate how different values for N influence the size of empirical confidence region (calculated from the model fittings to the artificial data), and consequently how many points are included.

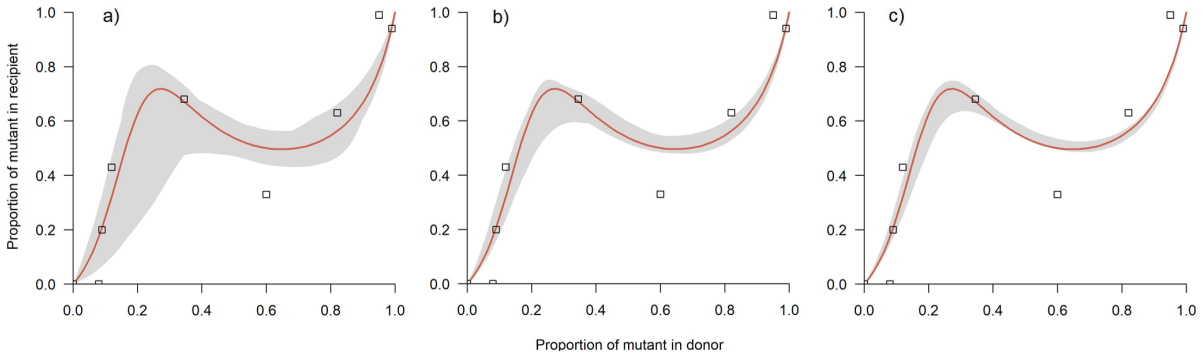

Figure 4.9: Empirical confidence regions from model fitting to artificial data for different values of N: a) $N = 60$, b) $N = 225$, c) $N = 445$.

This new criterion was studied through the use of simulations, similar to the procedure

described in the previous section. With N being fixed sequentially in the range $[5, 500]$, we generate artificial data, $\hat{p}_r(N, \hat{\boldsymbol{\theta}})$. For each N, we repeat 20 times and compute the proportion of data points that are included within 95% of the artificial data. Then we rank the different values of N based on how each value of N fulfils this criterion.

As expected, if this filtering criterion is based on the amount of data points covered by the confidence intervals obtained, it will inevitably favour low values of N. Lower values of N create more stochasticity in the data, which will result then in wider empirical confidence regions.

Figure 4.10 shows the results of 100 simulations using only this criterion, suggesting the best estimate for $N = 25$, which is responsible for a higher proportion of data points covered via N-induced stochasticity.

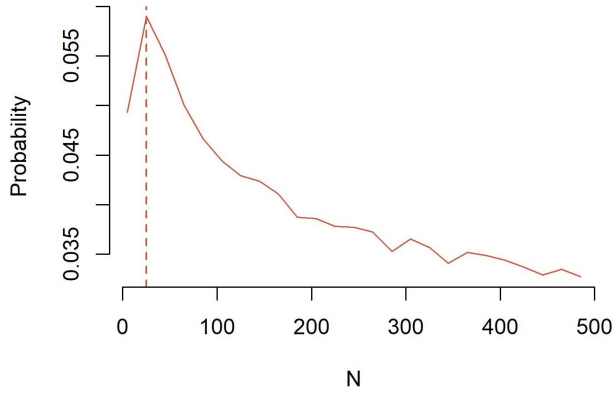

Figure 4.10: Ranking of N, based on the data point coverage criterion. Choosing to cover more points with the resulting confidence region, biases the best estimates of N to low values, giving more weight to stochasticity.

4.3.3 Estimation of N integrating both criteria

Thus as we have seen, there is a trade-off: small N induces more variability and is likely to capture more points, but its mean will be farther from the data (larger D), whereas large N, leads to less variability and is prone to capture less points, but its mean stochastic realizations will be on average closer to the data (smaller D). We illustrate these two opposing forces in Figure 4.11 through the use of simulations.

Combining the two filtering criteria on our $(N, \hat{\boldsymbol{\theta}})$-generated data leads to an intermediate range for N (see Figure 4.12) being most appropriate to capture the data points, while preserving sufficient accuracy.

Finally, using this distribution for N in stochastic simulations, with a bottleneck at time $\tau = 0$, and fixed $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}$, we obtain N from the filtering procedure. We use such simulations as

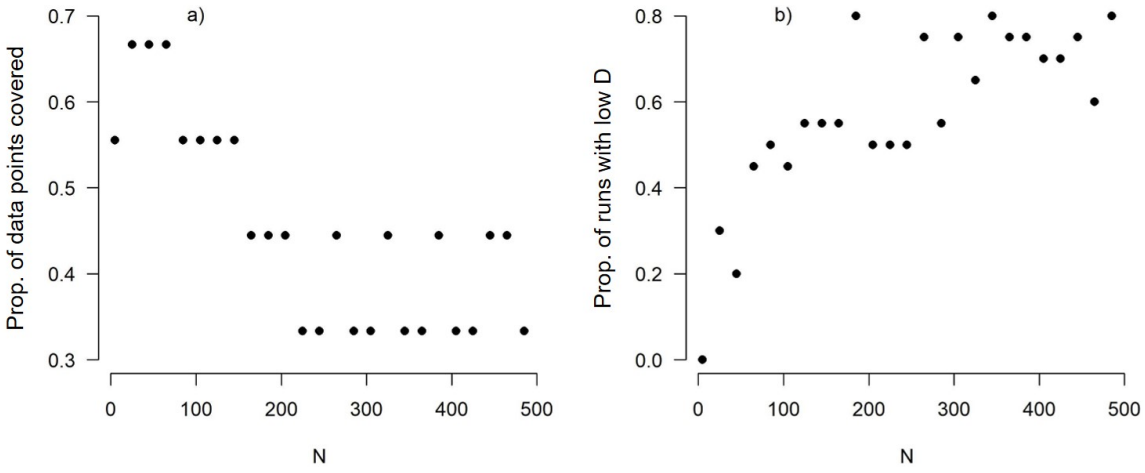

Figure 4.11: Trade-off between data point capture and quality of fit in $(N, \hat{\boldsymbol{\theta}})$-simulated data. a) As the value of N increases, the empirical confidence region obtained from the simulation approach is smaller so the proportion of data points covered is smaller. b) However, as N increases, the effect of stochasticity decreases, so the proportion of iterations close to the data points, i.e. that satisfy condition 4.3, is higher. By combining the two filtering criteria, we expect an intermediate value for N being optimal.

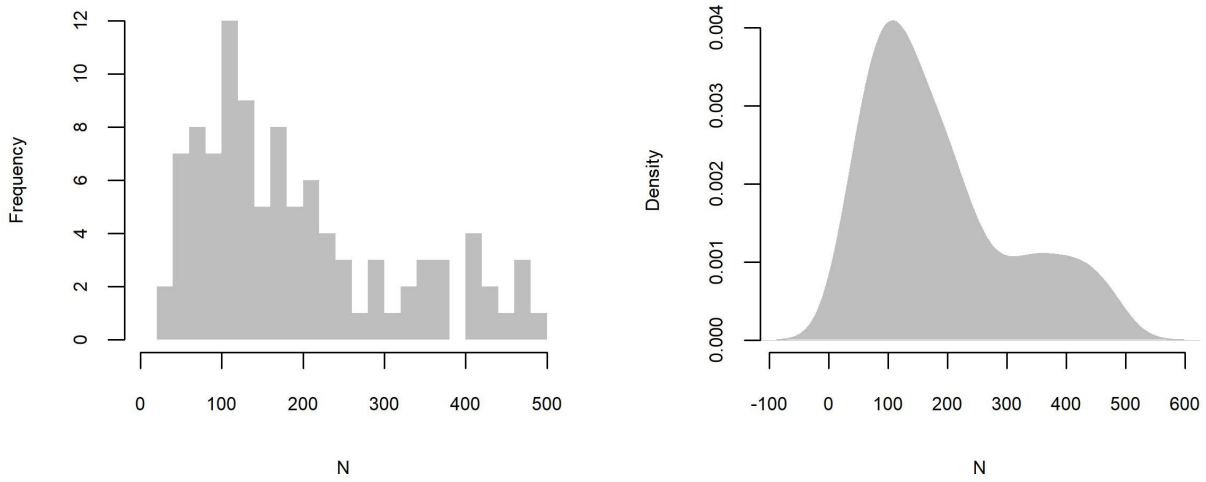

Figure 4.12: Empirical distribution of N from the $(N, \hat{\boldsymbol{\theta}})$-generated data filtering combining the data distance and data coverage criteria. a) Histogram of the values of N obtained from the simulation approach and b) corresponding smoothed distribution using `density` in R.

new artificial data and refit the model to obtain different parameter estimates. We accumulate all such parameter estimates and model fits and construct the 95% confidence region (Figure 4.13) and inferred parameter estimates (Table 4.4). For a study of parameter depencies, see Supplemental Section S3.

4.4 Conclusions

To account for uncertainty from our deterministic predictions, we applied two different methods. These considered different sources of stochasticity: experimental error or bottleneck effect. Both

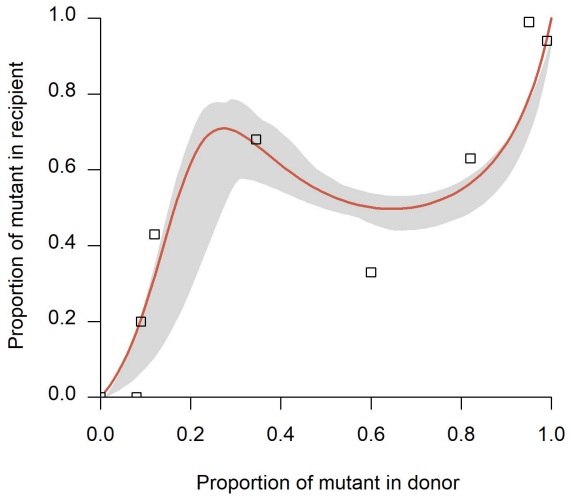

Figure 4.13: Model fitting to data and empirical confidence region obtained from simulations. The distribution of N derived previously (Figure 4.12) was used to simulate data to take into account the bottleneck size during transmission.

Table 4.4: Model parameter estimates from the model fitting to the H274Y data and parameter 95% confidence intervals obtained from simulating data based on the bottleneck effect. The estimates presented for N are the mean and 95% quantiles from the distribtuion of N shown in 4.12.

$\boldsymbol{\theta}$	$\hat{\boldsymbol{\theta}}$	95% CI
ρ	96.325	(96.325, 96.326)
a_{12}	0.759	(0.619, 0.776)
a_{21}	0.951	(0.933, 0.958)
β	0.999	(0.997, 1.012)
N	232	(26, 474)

are consistent with low deviation around the original data and the best-fitting model parameters, but the last one is more informative since it connects more variability in p_d to variability in p_r and mechanistically to the number of virions transmitted. After both uncertainty quantification, we were able to estimate uncertainty around the parameters that is not too big, and an estimate for N that reflects an intermediate number of virions upon transmission. We estimated a mean bottleneck of $N \approx 230$. This is much higher than the estimate of McCaw et al. (2011) et al. ($N \approx 4$). Yet our estimate is consistent with the recent influenza literature ($N \approx 196$ in (Leonard et al., 2017) and $100 < N < 200$ in (Poon et al., 2016), both from influenza transmission data in humans). Thus, our framework provides a reasonable and plausible alternative for modeling and interpreting influenza strain mixture experiments.

Chapter 5

Discussion

The mathematical modeling of pathogen dynamics at multiple levels (within-host and between-hosts) is of crucial importance to gain insights about the infection treatment. Computational models have been essential to guide us about the costs and benefits of intervention strategies, and to give us an understanding of the ecological and evolutionary dynamics of pathogens in general, and of influenza viruses in particular.

Competition is an integral ecological interaction and viruses compete for the limited host cells. In our model, competition, although modeled in a simple form, plays a central and explicit role, and can inform us which strain prevails at an epidemiological scale. We have shown how we can apply a relatively simple within-host model to transmission data. In this setup, the parameter estimates not only give information about dynamics in one host, but also a deeper understanding of which viral strain could be more prevalent in the population after several rounds of such transmission events.

From the model validation studies, we can recover close parameter estimates to the true simulated parameters. This means we could confidently use this model fitting procedure to infer relvant parameter estimates. When applied to the H274Y data of McCaw et al. (2011), the model then predicts an approximately 95 times higher rate of growth of the wild-type strain in comparison to the mutant strain, contrary to the estimates of McCaw et al. (2011). We also infer, however, a relatively higher intra-strain competition of the wild-type in comparison to the mutant. Our model predicts a scenario of coexistence, given that a_{12} and a_{21} are both smaller than 1. However, the H274Y data was limited, so although this work is a step forward in terms of proposing modeling alternatives, it should also be taken as a call for better data to probe more complex scenarios of fitness differences and outcomes between pathogen strains.

Another experimental recommendation is the use of more relative strain proportions closer to 50/50. Proportions near absolute value for any of the two strains are not informative since for any scenario and for any initial conditions, the system tends to either one strain leading the other to extinction or vice-versa. Additionally, the availability of total number of virions, not just proportions, as quantities measured experimentally would be more informative for the entire explicit parametrization of the system. As we considered only proportions, we could only estimate "relative ratios" between within-host growth and competition parameters. If the total population size is also modeled, we would be able to quantify the transmission and competition dynamics more in detail.

Our proposed framework has implications at the epidemiological level. If scenario 1, where the mutant strain outcompetes the wild-type, is the most prevalent at the within-host level, it is expected that the mutant strain is the prevailing strain at the population level. The opposite would be anticipated if the second scenario was the more common. However, if scenario 3, where strain co-existence is achieved, both could be transmitted between hosts, increasing the heterogeneity at the population level and making, for example, vaccine design more challenging.

Viruses must transmit to new susceptible hosts in order to replicate, but not all viral particles get transferred. A central feature of the transmission is the bottleneck, i.e. how many virions from the donor enter the recipient. Many factors may be responsible for the transmission bottleneck, from different immunological conditions in the host to simple intrinsic stochastic effects. Viruses are subject to different types of bottlenecks: during transmission, during organ/tissue colonization and during cell infection (also referred as multiplicity of cellular infection - MOI) (Gutiérrez et al., 2012). However, in the context of this thesis, we have modeled only the effect of the transmission bottleneck.

The total number of virions transmitted, N, has a relevant impact in a mixture transmission since it increases the sampling error the smaller it is, possibly masking the true fitness hierarchies between strains. This means that if fewer total virions are transmitted between hosts, the less relevant becomes the relative strain proportions in the donor because the outcome is more stochastic and unpredictable.

From our simulation method, we estimate a mean bottleneck size of $N \approx 230$, which is consistent with recent quantification studies on influenza (Leonard et al., 2017; Poon et al., 2016). It is however much higher than the 3.8 virions proposed by McCaw et al. It has been proposed that a possible explanation for this is that infection in ferrets may need less influenza viral particles (Sigal et al., 2018) or that the competitive-mixtures experimental method is subject

to high stochastic fluctuations in the recipient by looking only at two strains. Another study (McCrone et al., 2018) also estimated a very narrow bottleneck. This disparity in the transmission bottlenecks may be due to, among other possible explanations, the samples coming from climates with very different temperature and humidity (for example (Poon et al., 2016) used samples from subtropical climate while those from (McCrone et al., 2018) were from a temperate climate), and this is know to affect influenza transmission (Lowen et al., 2007).

Another caveat of our model is the absence of any explicit immunity effects. At the onset of an influenza infection, innate immunity is activated, and adaptive immunity starts at about 3 days post-infection (van de Sandt et al., 2012). Since we are focusing on transmission events, a reasonable assumption is that the observed viral dynamics in the recipient hosts happen quickly. In other words we make a quasi steady state (QSS) assumption, where the predicted dynamics happen at an early time frame, before any immune activation. However, our assumption of neglecting the acute phase of infection, is similar to the assumption in the model by McCaw et al. (2011), where exponential growth was assumed in the time-frame of the experiment.

Our study has implications for epidemiology, mathematical modeling and for understanding experimental results in competitive-mixture designs in general. Allowing for the possibility of frequency-dependent competition and hierarchies between strains, it expands the range of possible scenarios that can be captured with such model, including coexistence and bistability, prior to immune activation. Furthermore, with mutual coexistence as a plausible outcome for the competition dynamics between two strains at the within-host level, the population-level coexistence becomes even easier to explain, taking into account their simultaneous co-transmission from host to host. Interestingly, the other scenario of bistability within-host suggests that depending on initial frequencies and stochasticity upon initial contact, one strain or the other may win at the within-host level. These two ways of coexistence suggest high within-host diversity and low between-host diversity in one case (within-host coexistence) and low within-host diversity coupled with high between-host diversity in the other (bistability within host). Correctly quantifying and disentangling these two scenarios of maintenance of pathogen diversity may be of paramount importance when designing control strategies for different pathogens.

Furthermore, estimation of bottleneck size, another crucial quantity at the within-to between-host interface, is very tightly linked to the assumptions of the underlying model. The more flexible a model is to capture intricate non-linearity in the data, the less room there is for patterns to be assigned to stochasticity, for example to low bottleneck size. We based the estimation of the Lotka-Volterra model parameters just on the availability of proportion data at one snapshot in

time. This precludes the full identification of the 6 parameters of the explicit model. Under the availability of both proportion data and total viral count data over more time points, we expect the full model parametrization to be possible. Naturally, separating signal from noise remains a challenging problem in all areas of parameter estimation and model fitting to data, but when using a more complex model and when we have high confidence in the quality of the data, we can test for more complex biological signal and examine more refined hypotheses. Thus, although this work is a step forward in terms of proposing alternatives to modeling, it should also be taken as a call for better data to probe more complex scenarios of fitness differences and outcomes between pathogen strains.

Bibliography

P. Baccam, C. Beauchemin, C. A. Macken, F. G. Hayden, and A. S. Perelson. Kinetics of influenza A virus infection in humans. *Journal of virology*, 80(15):7590–7599, 2006.

C. A. Beauchemin and A. Handel. A review of mathematical models of influenza A infections within a host or cell culture: lessons learned and challenges ahead. *BMC public health*, 11(1): S7, 2011.

G. Bocharov and A. Romanyukha. Mathematical model of antiviral immune response III. influenza A virus infection. *Journal of Theoretical Biology*, 167(4):323–360, 1994.

A. Boianelli, V. Nguyen, T. Ebensen, K. Schulze, E. Wilk, N. Sharma, S. Stegemann-Koniszewski, D. Bruder, F. Toapanta, C. Guzmán, et al. Modeling influenza virus infection: a roadmap for influenza research. *Viruses*, 7(10):5274–5304, 2015.

T. Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

R. H. Byrd, P. Lu, J. Nocedal, and C. Zhu. A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, 16(5):1190–1208, 1995.

Y.-H. Cheng, S.-H. You, Y.-J. Lin, S.-C. Chen, W.-Y. Chen, W.-C. Chou, N.-H. Hsieh, and C.-M. Liao. Mathematical modeling of postcoinfection with influenza A virus and Streptococcus pneumoniae, with implications for pneumonia and copd-risk assessment. *International journal of chronic obstructive pulmonary disease*, 12:1973, 2017.

B. J. Coburn, B. G. Wagner, and S. Blower. Modeling influenza epidemics and pandemics: insights into the future of swine flu (H1N1). *BMC medicine*, 7(1):30, 2009.

E. Domingo. Mechanisms of viral emergence. *Veterinary research*, 41(6):38, 2010.

S. Duan, D. A. Boltz, P. Seiler, J. Li, K. Bragstad, L. P. Nielsen, R. J. Webby, R. G. Webster, and E. A. Govorkova. Oseltamivir–resistant pandemic H1N1/2009 influenza virus possesses lower transmissibility and fitness in ferrets. *PLoS pathogens*, 6(7):e1001022, 2010.

Z. Feng, S. Towers, and Y. Yang. Modeling the effects of vaccination and treatment on pandemic influenza. *The AAPS journal*, 13(3):427–437, 2011.

N. M. Ferguson, D. A. Cummings, S. Cauchemez, C. Fraser, S. Riley, A. Meeyai, S. Iamsirithaworn, and D. S. Burke. Strategies for containing an emerging influenza pandemic in Southeast Asia. *Nature*, 437(7056):209, 2005.

F. Fröhlich, C. Loos, and J. Hasenauer. Scalable inference of ordinary differential equation models of biochemical processes. In *Gene Regulatory Networks*, pages 385–422. Springer, 2019.

T. C. Germann, K. Kadau, I. M. Longini, and C. A. Macken. Mitigation strategies for pandemic influenza in the united states. *Proceedings of the National Academy of Sciences*, 103(15): 5935–5940, 2006.

M. Gillman. *An introduction to mathematical models in ecology and evolution: time and space*, volume 4. John Wiley & Sons, 2009.

S. Gutiérrez, Y. Michalakis, and S. Blanc. Virus population bottlenecks during within-host progression and host-to-host transmission. *Current opinion in virology*, 2(5):546–555, 2012.

B. Hancioglu, D. Swigon, and G. Clermont. A dynamical model of human immune response to influenza A virus infection. *Journal of theoretical biology*, 246(1):70–86, 2007.

B. P. Holder and C. A. Beauchemin. Exploring the effect of biological delays in kinetic models of influenza within a host or cell culture. *BMC Public Health*, 11(1):S10, 2011.

A. C. Hurt, S. S. Nor'e, J. M. McCaw, H. R. Fryer, J. Mosse, A. R. McLean, and I. G. Barr. Assessing the viral fitness of oseltamivir-resistant influenza viruses in ferrets, using a competitive-mixtures model. *Journal of virology*, 84(18):9427–9438, 2010.

N. H. Khanh. Stability analysis of an influenza virus model with disease resistance. *Journal of the Egyptian Mathematical Society*, 24:193–199, 2016.

B. Lee, L. Haidari, and M. Lee. Modelling during an emergency: the 2009 H1N1 influenza pandemic. *Clinical Microbiology and Infection*, 19(11):1014–1022, 2013.

A. S. Leonard, D. B. Weissman, B. Greenbaum, E. Ghedin, and K. Koelle. Transmission bottleneck size estimation from pathogen deep-sequencing data, with an application to human influenza A virus. *Journal of virology*, 91(14):e00171–17, 2017.

A. C. Lowen, S. Mubareka, J. Steel, and P. Palese. Influenza virus transmission is dependent on relative humidity and temperature. *PLoS pathogens*, 3(10):e151, 2007.

B. D. Martin and E. Schwab. Current usage of symbiosis and associated terminology. *International Journal of Biology*, 5(1):32, 2013.

J. M. McCaw, N. Arinaminpathy, A. C. Hurt, J. McVernon, and A. R. McLean. A mathematical framework for estimating pathogen transmission fitness and inoculum size using data from a competitive mixtures animal model. *PLoS Computational Biology*, 7(4):e1002026, 2011.

J. T. McCrone, R. J. Woods, E. T. Martin, R. E. Malosh, A. S. Monto, and A. S. Lauring. Stochastic processes constrain the within and between host evolution of influenza virus. *Elife*, 7:e35962, 2018.

J. McVernon, C. McCaw, and J. Mathews. Model answers or trivial pursuits? the role of mathematical models in influenza pandemic preparedness planning. *Influenza and other respiratory viruses*, 1(2):43–54, 2007.

R. Mikolajczyk, R. Krumkamp, R. Bornemann, A. Ahmad, M. Schwehm, and H.-P. Duerr. Influenza—insights from mathematical modelling. *Deutsches Ärzteblatt International*, 106 (47):777, 2009.

J. D. Murray. Mathematical Biology. Third edition, 2002.

K. Nicholson. Clinical features of influenza. In *Seminars in respiratory infections*, volume 7, pages 26–37, 1992.

C. R. Parrish and Y. Kawaoka. The origins of new pandemic viruses: the acquisition of new host ranges by canine parvovirus and influenza A viruses. *Annu. Rev. Microbiol.*, 59:553–586, 2005.

L. L. Poon, T. Song, R. Rosenfeld, X. Lin, M. B. Rogers, B. Zhou, R. Sebra, R. A. Halpin, Y. Guan, A. Twaddle, et al. Quantifying influenza virus diversity and transmission in humans. *Nature genetics*, 48(2):195, 2016.

C. W. Potter. A history of influenza. *Journal of applied microbiology*, 91(4):572–579, 2001.

P. Renard, A. Alcolea, and D. Gingsbourger. Stochastic versus deterministic approaches. In *Environmental Modelling: Finding Simplicity in Complexity, Second Edition (eds J. Wainwright and M. Mulligan)*, pages 133–149. Wiley Online Library, 2013.

P. Rodpothong and P. Auewarakul. Viral evolution and transmission effectiveness. *World journal of virology*, 1(5):131, 2012.

P. Saunders-Hastings, B. Q. Hayes, D. Krewski, et al. Modelling community-control strategies

to protect hospital resources during an influenza pandemic in Ottawa, Canada. *PloS one*, 12 (6):e0179315, 2017.

D. Sigal, J. N. Reid, and L. M. Wahl. Effects of transmission bottlenecks on the diversity of influenza A virus. *Genetics*, 210(3):1075–1088, 2018.

A. M. Smith. Host-pathogen kinetics during influenza infection and coinfection: insights from predictive modeling. *Immunological reviews*, 285(1):97–112, 2018.

J. C. Stack, P. R. Murcia, B. T. Grenfell, J. L. Wood, and E. C. Holmes. Inferring the inter-host transmission of influenza a virus using patterns of intra-host genetic variation. *Proceedings of the Royal Society B: Biological Sciences*, 280(1750):20122173, 2013.

Y. Suzuki. Sialobiology of influenza: molecular mechanism of host range variation of influenza viruses. *Biological and Pharmaceutical Bulletin*, 28(3):399–408, 2005.

S.-i. Tamura and T. Kurata. Defense mechanisms against influenza virus infection in the respiratory tract mucosa. *Jpn J Infect Dis*, 57(6):236–47, 2004.

D. A. Turner, A. J. Wailoo, K. G. Nicholson, N. Cooper, A. J. Sutton, and K. R. Abrams. Systematic review and economic decision modelling for the prevention and treatment of influenza A and B. *Health Technology Assessment*, 2003.

C. E. van de Sandt, J. H. Kreijtz, and G. F. Rimmelzwaan. Evasion of influenza a viruses from innate and adaptive immune responses. *Viruses*, 4(9):1438–1476, 2012.

A. R. Wargo and G. Kurath. Viral fitness: definitions, measurement, and current insights. *Current opinion in virology*, 2(5):538–545, 2012.

R. G. Webster. Antigenic variation in influenza viruses. In *Origin and evolution of viruses*, pages 377–390. Elsevier, 1999.

W. H. O. (WHO) et al. Influenza (seasonal) factsheet N 211. Geneva: WHO; 2009, 2019.

J. Wilson and M. Itzstein. Recent strategies in the search for new anti-influenza therapies. *Current drug targets*, 4(5):389–408, 2003.

J. T. Wootton and M. Emmerson. Measurement of interaction strength in nature. *Annu. Rev. Ecol. Evol. Syst.*, 36:419–444, 2005.

Y. Xu, L. J. Allen, and A. S. Perelson. Stochastic model of an influenza epidemic with drug resistance. *Journal of theoretical biology*, 248(1):179–193, 2007.

Supplemental material

S1. Optimization routine comparison

In this section we show the results of a small comparison study of optimization functions. The chosen functions can be applied to non-linear problems, including ODE systems. The functions differ in their convergence criteria so the estimates may diverge. They were subject to a single data fitting routine, all with the same fixed initial guess for the parameters. Table 5.1 summarizes the conditions and results of the fitting procedures used to compare the functions.

Table 5.1: Data fitting results of different optimization functions.

Function	$\boldsymbol{\theta}$ constrains	$\boldsymbol{\theta}$ estimates				MSE	Time
		ρ	a_{12}	a_{21}	β		
optim	$[0, \infty[$	95.514	0.754	0.949	0.999	0.0129	7.92
optimx	$[0, \infty[$	95.514	0.754	0.949	0.999	0.0129	8.66
nlminb	$[0, \infty[$	853.203	0.975	0.995	0.999	0.0125	11.25
nlm	–	0.050	0.007	-25.096	0.062	0.0215	25.53
solnl	$[0, \infty[$	1.619	0.000	0.054	0.736	0.0223	0.74

The function `optimx` is an extension of `optim`. It converges to the same estimates, and minimizes the MSE to same value, however it takes slightly longer. The time difference may be small, but at larger scales like in simulations, it becomes significant. `nlm` cannot be applied to our system since it does not accept box constrains, i.e. lower or upper bounds for the parameters, and they must be positive in our model. `solnl` is the fastest because it runs less iterations, at the expense of not minimizing the MSE as much as other optimization functions. This puts even more weight in the initial guesses that we fix for the parameters, which should be avoided. Therefore, `optim` was a safe choice to apply to our system.

S2. Local and global optima assessment

Like mentioned in Chapter 3, an optimization may not always converge to the global minimum if given inappropriate guesses for the parameters. We ran 100 simulations, each with a different set of initial guesses for the parameters, to test for possible non-identifiability.

We show in Table 5.2 the parameter estimates from the previous simulations, ordered by ascending values of MSE. As we can see, the values with higher quality-of-fit, i.e. smaller MSE, have parameter estimates very close to the ones estimated from the model fitting in Section 4.1.

Table 5.2: Simulation results of assessment of the $\boldsymbol{\theta}$ guess choice effect. 100 simulations with different starting parameter guesses (ρ_0, a_{12_0}, a_{21_0}, β_0) were carried. Those with smaller MSE values have the $\boldsymbol{\theta}$ estimates closer to $\hat{\boldsymbol{\theta}}$ obtained in Section 4.1, so we can conclude that by choosing model fits with the least error we guarantee parameter identifiability.

ρ_0	a_{12_0}	a_{21_0}	β_0	$\hat{\rho}$	$\hat{a}_{12}$	$\hat{a}_{21}$	$\hat{\beta}$	MSE
14.581	0.918	0.426	0.966	94.244	0.753	0.949	0.999	0.0129
9.868	0.940	0.725	0.924	90.858	0.746	0.746	0.999	0.0128
6.891	0.584	0.149	0.716	90.858	0.746	0.948	0.999	0.0127
86.936	0.010	0.577	0.730	90.858	0.746	0.948	0.999	0.0127
18.797	0.249	0.420	0.317	90.858	0.746	0.948	0.999	0.0127

If we filter the simulations based on MSE (MSE < 0.013, in this case), we obtain distributions of the parameters that peak at their best estimates, $\hat{\boldsymbol{\theta}}$ (Figure 5.1).

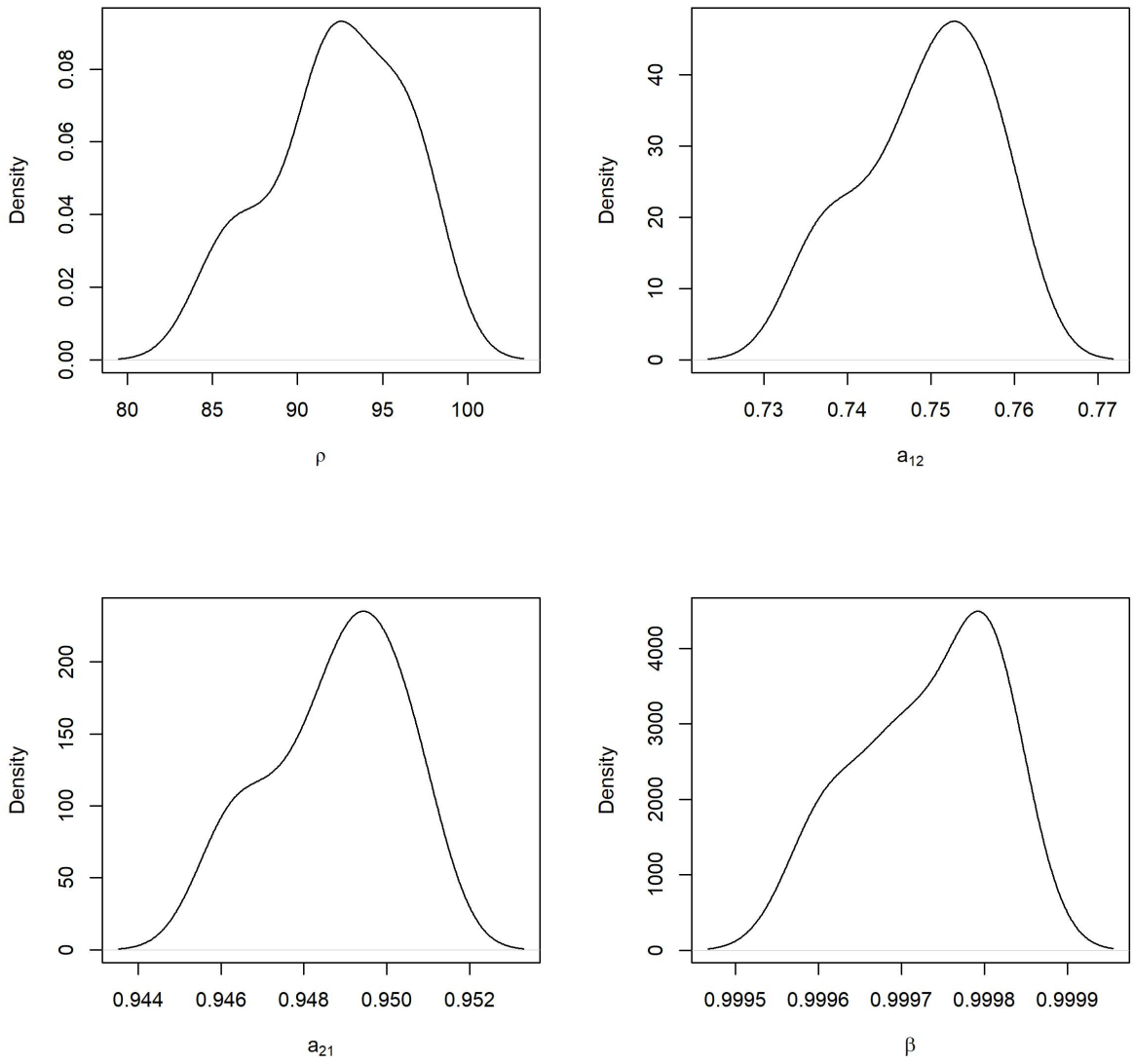

Figure 5.1: Assessment of non-uniqueness in parameter estimation. We filter the 100 simulations based on MSE and find that all empirical parameter distributions peak at their best estimates.

S3. Parameter dependencies

Figure 5.2 shows the combinations of the estimated parameters obtained from 100 simulations, without any filtering by MSE. This serves as a visual assessment of possible inter dependencies between model parameters. If correlations between parameters are observed, the estimation of some parameters may influence the estimation of others and make optimization more difficult.

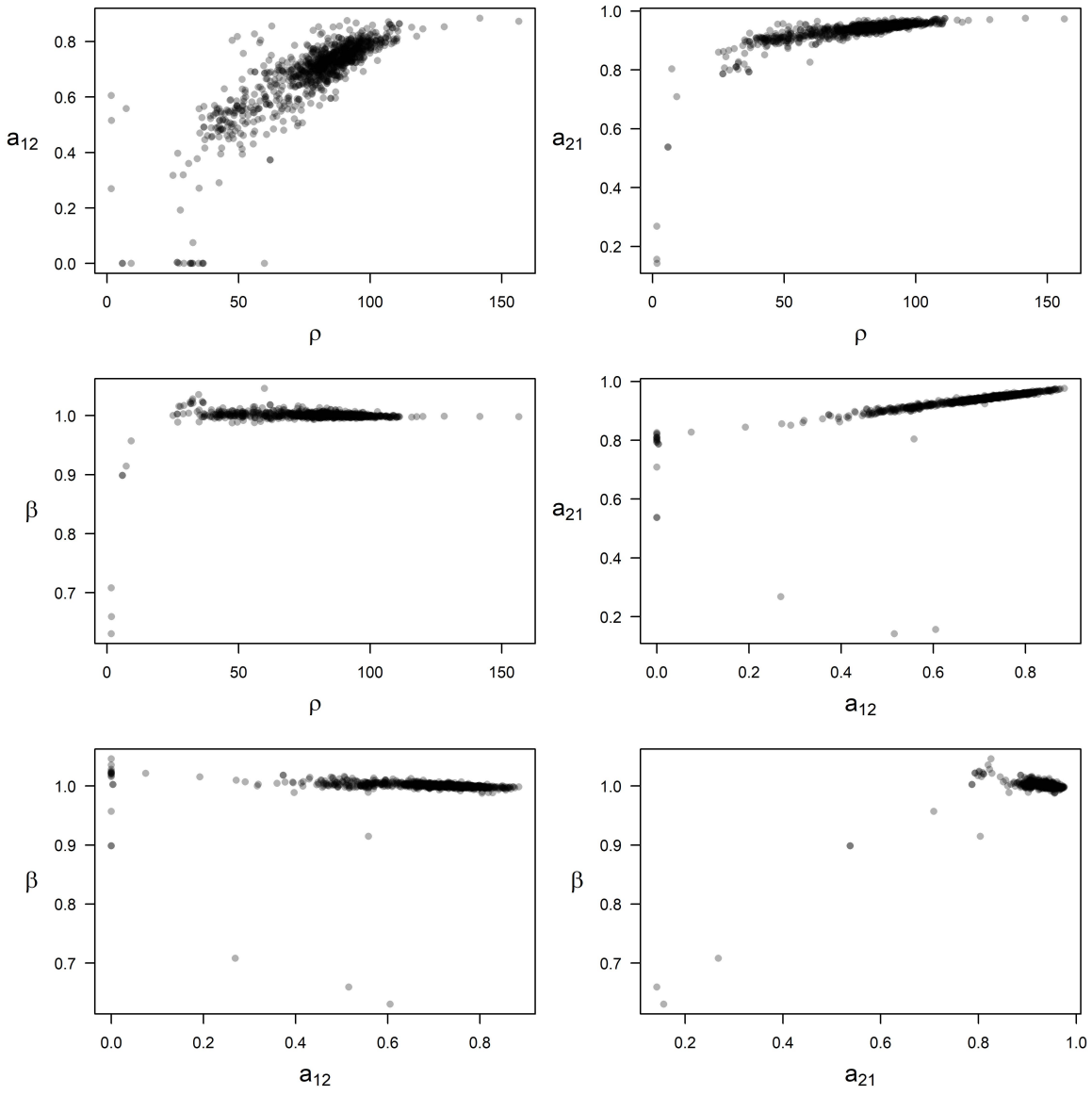

Figure 5.2: Scatter plots from the computed parameters values of the simulations.

It appears that the parameters ρ and a_{12} are correlated: increasing the value of ρ increases a_{12}.

S4. R code

Due to the large number different R scripts used throughout this project, we present here only the main functions applied to the rescaled model. The codes depend on the library deSolve.

Initial conditions and data.

```
Tpoint <- 1; times <- 0:1; nsteps <- 9
p.donor <- c(0, 8, 9, 12, 34.5, 60, 82, 95, 99)/100
p.recei <- c(0, 0, 20, 43, 68, 33, 63, 99, 94)/100
```

Function that solves the ordinary differential equations.

```
pr.mod <- function(theta){
  p1SIM <- c()
  for(k in 1:nsteps){
    u1 <- p.donor[k]; beta <- theta[4]; u2 <- (1 - u1)/beta; ic <- c(u1 = u1, u2 = u2)
    difLVT <- function(t, ic, theta){
      with(as.list(c(ic, theta)),{
        du1dt <- u1*(1-u1-a12*u2)
        du2dt <- rho*u2*(1-u2-a21*u1)
        list(c(du1dt, du2dt))
      })
    }
    out <- ode(ic, times, difLVT, theta)
    p1sim <- unname(out[Tpoint+1, 2])/(unname(out[Tpoint+1, 2]) + unname(out[Tpoint+1, 3])*unname(
        beta))
    p1SIM <- c(p1SIM, p1sim)
  }
  return(p1SIM)
}
```

Same purpose as the previous function but includes the bottleneck effect.

```
pr.mod.CTS <- function(theta, N){
  p1SIM <- c()
  for(k in 1:length(p.donor)){
    u1 <- rbinom(1, N, p.donor[k])/N; beta <- theta[4]; u2 <- (1 - u1)/beta; ic <- c(u1 = u1, u2 =
        u2)
    difLVT <- function(t, ic, theta){
      with(as.list(c(ic, theta)),{
        du1dt <- u1*(1-u1-a12*u2)
        du2dt <- rho*u2*(1-u2-a21*u1)
```

```
        list(c(du1dt, du2dt))
      })
    }
    out <- ode(ic, times, difLVT, theta)
    p1sim <- unname(out[Tpoint+1, 2])/(unname(out[Tpoint+1, 2]) + unname(out[Tpoint+1, 3])*unname(
        beta))
    p1SIM <- c(p1SIM, p1sim)
  }
  return(p1SIM)
}
```

Function that calculates the MSE. This is the function to be minimized in the optimization.

```
calc.err <- function(theta){
  error <- pr.mod(theta) - data
  return(mean(error^2))
}
```

Data fitting. pFIT corresponds to the curve traced by the best parameter estimates (for that data and initial parameter guesses).

```
theta.guess <- c(rho = 1, a12 = 1, a21 = 1, beta = 1)
data <- p.recei
fit <- optim(par = theta.guess, fn = calc.err, method = "L-BFGS-B", lower = c(0,0,0,0), upper = c(
    Inf,Inf,Inf,Inf), hessian = T)
pFIT <- pr.mod(fit$par)
```

Simulation implementing the two criteria for filtering the artificial data when accounting for bottleneck-caused uncertainty.

```
allbestN <- vector()
distN <- vector()
bestTheta <- fit$par
for (bigruns in 1:100){
  Nvals <- seq(5, 500, by = 20) # range of N
  nruns <- 20  # stochastic realizations for each N
  probN <- vector()
  pipN <- vector()
  y <- matrix(nrow = length(Nvals), ncol = nruns)
  for (i in 1:length(Nvals)){
    N <- Nvals[i]
    z <- matrix(nrow = nruns, ncol = length(data.y))
```

```
    for (run in 1:nruns){
      y[i,run] <- mean((pr.modDIS.N(bestTheta, N)[1:9] - data.y)^2)
      z[run,] <- pr.modDIS.N(bestTheta, N)[1:9] # this stores the simulated data to check how many
          data points are covered
    }
    zquant1 <- apply(z, 2, quantile, probs = 0.025) # lower bound
    zquant2 <- apply(z, 2, quantile, probs = 0.975) # upper bound
    pipN[i] <- length(which(data.y > zquant1 & data.y < zquant2))/length(data.y) # how much data
        falls in between?
    countfreq <- length(which((y[i,] - bestMse)/bestMse < 0.2))
    probN[i] <- countfreq/nruns
  }
  distN <- rbind(distN, (pipN/max(pipN))^2*probN/max(probN)) # this product takes into account both
      criteria
  bestN <- Nvals[which(probN == max(probN, na.rm=T))]
  allbestN <- c(allbestN,bestN) # this allbestN takes into account only the mse criterion
  print(bigruns)
}
```